The American Economic Review

ARTICLES

MARCH 1969

THE AMERICAN ECONOMIC ASSOCIATION

Founded in 1885

• Published at George Banta Co., Inc., Menasha, Wisconsin.

• THE AMERICAN ECONOMIC REVIEW, including four quarterly numbers, the *Proceedings* of the annual meetings, the *Directory* and Supplements, is published by the American Economic Association and is sent to all members five times a year, in March, May, June, September and December.

• Membership dues of the Association are $10.00 a year, $8.00 of each year's dues being in payment of a year's subscription both to the *Review* and the *Journal of Economic Literature*. Subscriptions by nonmembers are $10.00 a year. Single copies of the *Review*, $2.00; of the *Journal*, $1.50.

• Correspondence relating to the *Paper and Proceedings*, the *Directory*, advertising, permission to quote, business matters, subscriptions, membership and changes of address may be sent to the secretary, Harold F. Williamson, 629 Noyes Street, Evanston, Illinois 60201. To be effective, notice of change of address must reach the secretary by the 1st of the month previous to the month of publication.

• Second-class postage paid at Menasha, Wisconsin. Printed in U.S.A.

THE AMERICAN ECONOMIC REVIEW

March 1969

VOLUME LIX, NUMBER 1

• Manuscripts and editorial correspondence relating to the regular quarterly issue of this REVIEW should be addressed to George H. Borts, Managing Editor of THE AMERICAN ECONOMIC REVIEW, Brown University, Providence, R.I. 02912. Manuscripts should be submitted in duplicate and in acceptable form. *Style Instructions* for guidance in preparing manuscripts will be provided upon request to the editor.

• No responsibility for the views expressed by authors in this REVIEW is assumed by the editors or the publishers, The American Economic Association.

Articles

Communications

EDITOR'S ANNOUNCEMENT:

At its March 1969 meeting, the Executive Board of The American Economic Association voted to discontinue payments to contributors of articles accepted after March 7, 1969.

We hope this action will not discourage the submission of manuscripts to the *Review*.

THE AMERICAN ECONOMIC REVIEW

June 1969

VOLUME LIX, NUMBER 3

Articles

• Manuscripts and editorial correspondence relating to the regular quarterly issue of this REVIEW should be addressed to George H. Borts, Managing Editor of THE AMERICAN ECONOMIC REVIEW, Brown University, Providence, R.I. 02912. Manuscripts should be submitted in duplicate and in acceptable form. *Style Instructions* for guidance in preparing manuscripts will be provided upon request to the editor.

Communications

THE AMERICAN ECONOMIC REVIEW

• Manuscripts and editorial correspondence relating to the regular quarterly issue of this REVIEW should be addressed to George H. Borts, Managing Editor of THE AMERICAN ECONOMIC REVIEW, Brown University, Providence, R.I. 02912. Manuscripts should be submitted in duplicate and in acceptable form. *Style Instructions* for guidance in preparing manuscripts will be provided upon request to the editor.

• No responsibility for the views expressed by authors in this REVIEW is assumed by the editors or the publishers, The American Economic Association.

December 1969
VOLUME LIX, NUMBER 5

Articles

Communications

Editor's Announcement:

As an experiment, the following paper from the Eighty-Second Annual Meeting of the American Economic Association has been included in this issue of the *Review*.

THE AMERICAN ECONOMIC REVIEW

VOLUME LIX

THE AMERICAN ECONOMIC ASSOCIATION

Executive Office: Evanston, Illinois
Editorial Office: Brown University, Providence, Rhode Island

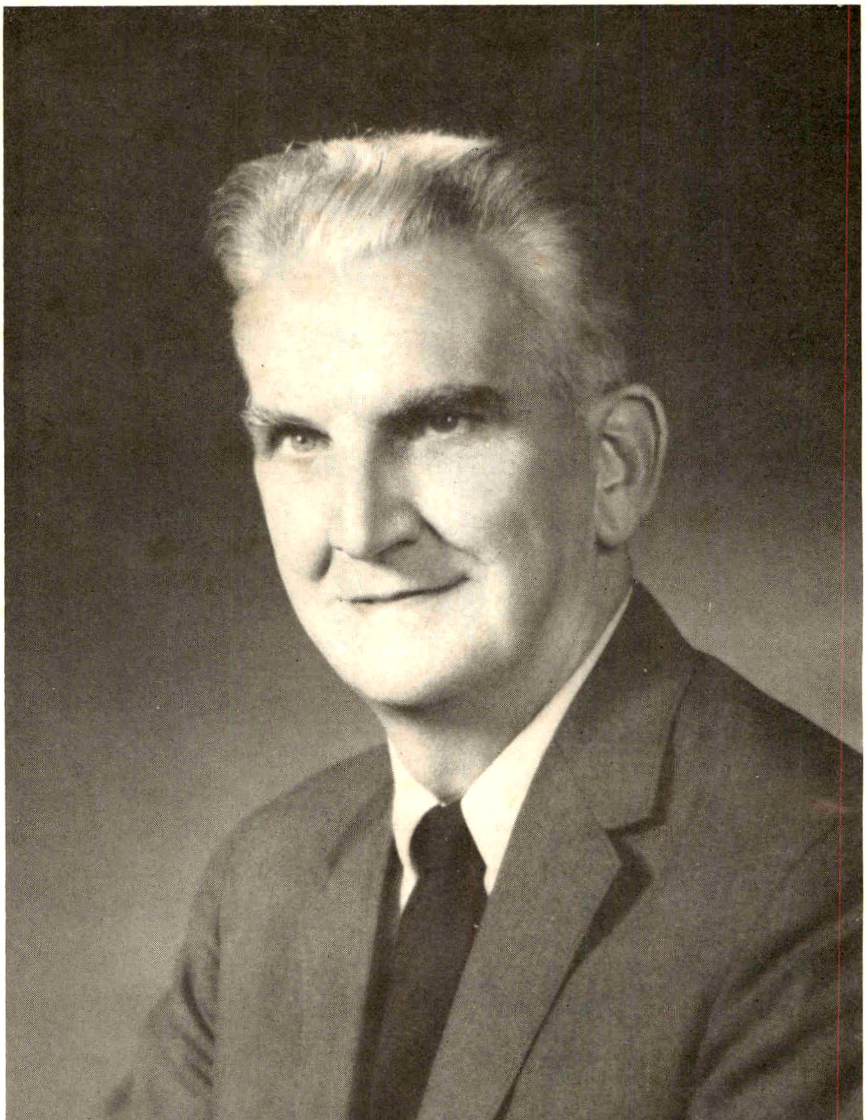
CONTENTS OF ARTICLES AND COMMUNICATIONS

CONTENTS OF PAPERS AND PROCEEDINGS

CONTRIBUTORS TO ARTICLES AND COMMUNICATIONS

CONTRIBUTORS TO PAPERS AND PROCEEDINGS

Number 70 of a series of photographs of past presidents of the Association.



K. E. Boulding

Economics As A Moral Science*

By Kenneth E. Boulding

Adam Smith, who has strong claim to being both the Adam and the Smith of systematic economics, was a professor of moral philosophy and it was at that forge that economics was made. Even when I was a student, economics was still part of the moral sciences tripos at Cambridge University. It can claim to be a moral science, therefore, from its origin, if for no other reason. Nevertheless, for many economists the very term "moral science" will seem like a contradiction. We are strongly imbued today with the view that science should be *wertfrei* and we believe that science has achieved its triumph precisely because it has escaped the swaddling clothes of moral judgment and has only been able to take off into the vast universe of the "is" by escaping from the treacherous launching pad of the "ought." Even economics, we learn in the history of thought, only became a science by escaping from the casuistry and moralizing of medieval thought. Who, indeed, would want to exchange the delicate rationality of the theory of equilibrium price, for the unoperational vaporings of a "just price" controversy? In the battle between mechanism and moralism generally mechanism has won hands down, and I shall not be surprised if the very title of my address does not arouse musty fears of sermonizing in the minds of many of my listeners.

Let me first explain, then, what I mean by moral and by moral science. A moral, or ethical proposition, is a statement about a rank order of preference among alternatives, which is intended to apply to more than one person. A preference which applies to one person only is a "taste." Statements of this kind are often called "value judgments." If someone says, "I prefer A to B," this is a personal value judgment, or a taste. If he says, "A is better than B," there is an implication that he expects other people to prefer A to B also, as well as himself. A moral proposition then is a "common value."

Every culture, or subculture, is defined by a set of common values, that is, generally agreed upon preferences. Without a core of common values a culture cannot exist, and we classify society into cultures and subcultures precisely because it is possible to identify groups who have common values.

Most tastes are in fact also common values and have been learned by the process by which all learning is done, that is, by mutation and selection. The most absurd of all pieces of ancient wisdom is surely the Latin tag *de gustibus non disputandum*. In fact, we spend most of our lives disputing about tastes. If we want to be finicky about definitions we might turn the old tag around and say where there is disputing, we are not talking about tastes. Nevertheless, even personal tastes are learned, in the matrix of a culture or a subculture in which we grow up, by very much the same kind of process by which we learn our common values. Purely personal tastes, indeed, can only survive in a culture which tolerates them, that is, which has a common value that private tastes of certain kinds should be allowed.

One of the most peculiar illusions of

* Presidential address delivered at the Eighty-first meeting of the American Economic Association, Chicago, Illinois, December 29, 1968. This address will be listed as Publication 122 of the Institute of Behavioral Science, University of Colorado.

economists is a doctrine that might be called the Immaculate Conception of the Indifference Curve, that is, that tastes are simply given, and that we cannot inquire into the process by which they are formed. This doctrine is literally "for the birds," whose tastes are largely created for them by their genetic structures, and can therefore be treated as a constant in the dynamics of bird societies. In human society, however, the genetic component of tastes is very small indeed. We start off with a liking for milk, warmth, and dryness and a dislike for being hungry, cold, and wet, and we do have certain latent drives which may guide the formation of later preferences in matters of sex, occupation, or politics, but by far and away the largest part of human preferences are learned, again by means of a mutation-selection process. It was, incidentally, Veblen's principal, and still largely unrecognized, contribution to formal economic theory, to point out that we cannot assume that tastes are given in any dynamic theory, in the sense that in dynamics we cannot afford to neglect the processes by which cultures are created and by which preferences are learned.

I am prepared indeed to go much further and to say that no science of any kind can be divorced from ethical considerations, as defined above. The propositions of science are no more immaculately conceived than the preferences of individuals. Science is a human learning process which arises in certain subcultures in human society and not in others, and a subculture as we have seen is a group of people defined by the acceptance of certain common values, that is, an ethic which permits extensive communication among them.

The scientific subculture is no exception to this rule. It is characterized by a strong common value system. A high value, for instance, is placed on veracity, on curiosity, on measurement, on quantification, on careful observation and experiment, and on objectivity. Without this common value structure the epistemological process of science would not have arisen; indeed it did not arise in certain societies where conditions might otherwise have been favorable but where some essential common values of the scientific subcultures did not exist. The question as to exactly what values and ethical propositions are essential to the scientific subculture may be in some dispute. The fact that there are such values cannot be disputed. It is indeed one of the most perplexing questions in intellectual history as to why the scientific subculture developed in the time and place that it did in Western Europe. The common values that are prerequisite to it are rather rare among human subcultures. The common values, for instance, of the military or the people that run the international system are quite different from those of science. In this sense, therefore, science has an essential ethical basis.

This means that even the epistemological content of science, that is, what scientists think they know, has an ethical component. The proposition, for instance, that water consists of two molecules of hydrogen and one of oxygen is not usually thought of as a proposition with high ethical content. Nevertheless, any student in chemistry who decides that he prefers to think of hydrogen as dephlogisticated water will soon find out that chemistry is not just a matter of personal taste. The fact that there is no dispute going on about any particular scientific proposition does not mean to say that it is a matter of taste; it simply means that the dispute about it has been resolved through the application of certain common values and ethical presuppositions.

There is however a fundamental sense in which the epistemological process even

in the physical and biological sciences is now running into situations which have strong ethical implications outside the scientific subculture. The myth that science is simply discovering knowledge about an objectively unchangeable world may have had some validity in the early stages of science but as the sciences develop this myth becomes less and less valid. The learning process of science is now running into two serious difficulties. The first might be called the generalized Heisenberg principle. When we are trying to obtain knowledge about a system by changing its inputs and outputs of information, these inputs and outputs will change the system itself, and under some circumstances they may change it radically. My favorite illustration of the Heisenberg principle is that of a man who inquires through the door of the bedroom where his friend is sick, "How are you?" whereupon the friend replies "Fine," and the effort kills him. In the social sciences of course the generalized Heisenberg principle predominates because knowledge of the social sciences is an essential part of the social system itself, hence objectivity in the sense of investigating a world which is unchanged by the investigation of it is an absurdity.

The second difficulty is that as science develops it no longer merely investigates the world; it creates the world which it is investigating. We see this even in the physical sciences where the evolution of the elements has now been resumed in this part of the universe after some six billion years. We are increasingly going to see this in the biological sciences, which will only find out about the evolutionary process by actively engaging in it, and changing its course. In the social sciences one can defend the proposition that most of what we can really know is what we create ourselves and that prediction in social systems can be achieved only by setting up consciously created systems which will make the predictions come true. Knowledge of random systems can only be obtained by destroying them, that is, by taking the randomness out of them. There is a great deal of evidence, for instance, that the fluctuations of prices in organized commodity or security markets are essentially random in nature. All we can possibly discover therefore by studying these fluctuations is what bias there might be in the dice. If we want to predict the future of prices in such a market we will have to control it, that is, we will have to set up a system of counterspeculation which will guarantee a given future course of prices. The gold standard is a primitive example of such a system in which it is possible to predict that the price of gold will lie within the gold points as long as the system remains intact. Similarly, we can predict the inside temperature of a house with an effective furnace and thermostat much better than we can predict the outside temperature simply because we control one and not the other.

We cannot escape the proposition that as science moves from pure knowledge toward control, that is, toward creating what it knows, what it creates becomes a problem of ethical choice, and will depend upon the common values of the societies in which the scientific subculture is embedded, as well as of the scientific subculture. Under these circumstances science cannot proceed at all without at least an implicit ethic, that is, a subculture with appropriate common values. The problem exists in theory even in what might be described as the objective phase of science, that is, the phase in which it is simply investigating "what is," because the question of the conditions under which ignorance is bliss is not an empty one. The assumption which is almost universal in academic circles that ignorance cannot possibly be bliss might under some circum-

stances be proved wrong by the very methods of science itself. As long as science is investigating an unchanging world, however, this problem does not become acute, for if knowledge does not change the world, then all ignorance does for us is to prevent us from satisfying our idle curiosity. When, however, knowledge changes the world the question of the content of the common values, both of the subculture which is producing knowledge and of the total society in which that subculture is embedded, becomes of acute importance. Under these circumstances the concept of a value-free science is absurd, for the whole future of science may well rest in our ability to resolve the ethical conflicts which the growth of knowledge is now creating. Science *could* create an ethical dynamic which would bring it to an end.

Let us return then to economics as a moral science, not merely in the sense in which all science is "affected with an ethical interest," but in the quite specific sense of asking whether economics itself can be of assistance in resolving ethical disputes, especially those which arise out of the continued increase of knowledge.

Economics specializes in the study of that part of the total social system which is organized through exchange and which deals with exchangeables. This to my mind is a better definition of economics than those which define it as relating to scarcity or allocation, for the allocation of scarce resources is a universal problem which applies to political decisions and political structures through coercion, threat, and even to love and community, just as it does to exchange. I have elsewhere distinguished three groups of social organizers which I have called the threat system, the exchange system, and the integrative system. Economics clearly occupies the middle one of these three. It edges over towards the integrative system insofar as it has some jurisdiction over the study of the system of one-way transfers of exchangeables, which I have called the "grants economy," for the grant, or one-way transfer, is a rough measure of an integrative relationship. On the other side, economics edges towards an area between the threat system and the exchange system which might be described as the study of strategy or bargaining.

To complete the circle there is also an area, between the threat system and the integrative system, of legitimated threat which is the principal organizer of political activity and the main subject matter of political science. All these systems are linked together dynamically through the process of human learning which is the main dynamic factor in all social systems. Part of this learning process is the learning of common values and moral choices, without which no culture and no social system is possible. The process by which we learn our preference structures indeed is a fundamental key to the total dynamics of society.

Economics, as such, does not contribute very much to the formal study of human learning, though some philosophical economists like Frederick Hayek [4] have made some interesting contributions to this subject. Our main contribution as economists is in the description of what is learned; the preference functions which embody what is learned in regard to values, and the production functions which describe the results of the learning of technology. We may not have thought much about the genetics of knowledge, but we have thought about its description, and this is a contribution not to be despised.

Thus, economics suggests the proposition that actual choices depend not only on preferences but on opportunities, and that under some circumstances quite small changes in either preferences or opportu-

nities may result in large changes in actual choices made. This proposition applies just as much to ethical choices and common values as it does to private tastes. It throws a good deal of light also on what might be called the evolutionary ecology of ethical systems. Successful ethical systems tend to create subcultures, and these subcultures tend to perpetuate and propagate the ethical systems which created them. This principle helps to explain the persistent division of mankind into sects, nations, and ideological groups. If we were to map the ethical preference systems of the individuals who comprise mankind, we would not find a uniform distribution but we would find a very sharp clustering into cultures and subcultures with relatively empty spaces between the clusters. All the members of a single sect, for instance, tend to think rather alike in matters of ethical judgment and differentiate themselves sharply from the ethical judgments of other sects. Individuals tend to be attracted to one or another of these clusters, leaving the social space between them relatively empty, like space between the stars. The reasons for this phenomenon lie deep in the dynamics of the human learning process, for our preferences are learned mainly from those with whom we have the most communication. This principle accounts for the perpetuation of such clusters, though it does not necessarily account for their original formation, which exhibits many puzzling phenomena. The splitting of these clusters in a kind of mitosis is also an important and very puzzling phenomenon. Once we realize, however, that these are highly sensitive systems as economic analysis suggests, we can see how wide divergences might arise. Thus, the actual difference in preferences and even opportunities between, shall we say, the socialist countries and the capitalist countries, may in fact be quite small, but this difference is enough to produce a very wide difference in the choices made.

Economics has made its own attempt to solve some of the problems involved in the moral judgment in what we know as welfare economics. I believe this attempt has been a failure, though a reasonably glorious one, and we should take a brief look at it. Welfare economics attempts to ask the question "What do we mean when we say that one state of a social system is better than another in strictly economic terms?" The most celebrated answer given is the Paretian optimum, which states in effect that Condition A of a social system is economically superior to Condition B, if nobody feels worse off in A than in B, and if at least one person feels better off. "Better off" or "worse off" are measured of course by preferences, so that we could restate the condition as saying that State A is superior to State B if one or more persons prefer A and if nobody prefers B. If we permit internal redistributions within the system, that is, compensation, the range of possible superior states is considerably broadened. From this simple principle a wide range of applications has been found possible in a stirring intellectual drama which might well be subtitled "Snow White (the fairest of all) and the Seven Marginal Conditions."

Many, if not most, economists accept the Paretian optimum as almost self-evident. Nevertheless, it rests on an extremely shaky foundation of ethical propositions. The more one examines it, for instance, the more clear it becomes that economists must be extraordinarily nice people even to have thought of such a thing, for it implies that there is no malevolence anywhere in the system. It implies, likewise, that there is no benevolence, the niceness of economists not quite extending as far as good will. It assumes selfishness, that is, the independence of in-

dividual preference functions, such that it makes no difference to me whether I perceive you as either better off or worse off. Anything less descriptive of the human condition could hardly be imagined. The plain fact is that our lives are dominated by precisely this interdependence of utility functions which the Paretian optimum denies. Selfishness, or indifference to the welfare of others, is a knife edge between benevolence on the one side and malevolence on the other. It is something that is very rare. We may feel indifferent towards those whom we do not know, with whom we have no relationships of any kind, but towards those with whom we have relationships, even the frigid relationship of exchange, we are apt to be either benevolent or malevolent. We either rejoice when they rejoice, or we rejoice when they mourn.

The almost complete neglect by economists of the concepts of malevolence and benevolence cannot be explained by their inability to handle these concepts with their usual tools. There are no mathematical or conceptual difficulties involved in inter-relating utility functions, provided that we note that it is the perceptions that matter [2]. The familiar tools of our trade, the indifference map, the Edgeworth box, and so on, can easily be expanded to include benevolence or malevolence, and indeed without this expansion many phenomena, such as one-way transfers, cannot be explained. Perhaps the main explanation of our neglect of these concepts is the fact that we have concentrated so heavily on exchange as the object of our study, and exchange frequently takes place under conditions of at least relative indifference or selfishness, though I argue that there is a minimum degree of benevolence even in exchange without which it cannot be legitimated and cannot operate as a social organizer. We exchange courtesies, smiles, the time of day and so on with the clerk in the store, as well as exchanging money for commodities. The amount of benevolence which exchangers feel towards each other need not be large, but a certain minimum is essential. If exchangers begin to feel malevolent toward each other exchange tends to break down, or can only be legitimated under conditions of special ritual, such as silent trade or collective bargaining.

Nevertheless, economists can perhaps be excused for abstracting from benevolence and malevolence, simply because their peculiar baby, which is exchange, tends to be that social organizer which lies between these two extremes, and which produces, if not selfishness, at least low levels of malevolence and benevolence. The threat system constantly tends to produce malevolence simply because of the learning process which it engenders. A threatener may begin by feeling benevolent toward the threatened—"I am doing this for your own good"—but threats almost invariably tend to produce malevolence on the part of the threatened towards the threatener, and this is likely to produce a type of behavior which will in turn produce malevolence on the part of the threatener towards the threatened. This can easily result in a cumulative process of increasing malevolence which may or may not reach some kind of equilibrium. The breakup of communities into factions and into internal strife frequently follows this pattern. At the other end of the scale, the integrative system tends to produce benevolence and those institutions which are specialized in the integrative system, such as the family, the church, the lodge, the club, the alumni association, and so on, tend also to create and organize benevolence, even beyond the circle of their members. This is partly because benevolence seems to be an important element in establishing a satisfactory personal identity, especially after the

threat system has been softened by the development of exchange. Those who live under threat, who generally occupy the lower end of the social scale, as well as those who live by threat at the upper end, tend to find their personal identities through malevolence and through the development of counter-threat or through the displacement of hatred onto weaker objects, such as children and animals. Once this state is passed, however, and society is mainly organized by exchange, there seems to be a strong tendency to move towards the integrative system and the integrative institutions. The Rotary Club is a logical extension of a business-oriented society, but it is not one that would necessarily have occurred to economists.

Oddly enough, it is not welfare economics with its elegant casuistry, subtle distinctions, and its ultimately rather implausible recommendations, which has made the greatest impact on the development of common values and ethical propositions. The major impact of economics on ethics, it can be argued, has come because it has developed broad, aggregative concepts of general welfare which are subject to quantification. We can see this process going right back to Adam Smith, where the idea of what we would today call per capita real income, as the principal measure of national well-being, has made a profound impact on subsequent thinking and policy. The development of the concept of a gross national product and its various modifications and components as statistical measures of economic success, likewise, has had a great impact in creating common values for the objectives of economic policy. Another, less fortunate, example of a measure which profoundly affected economic policy was the development of the parity index by the Bureau of Agricultural Economics in the United States Department of Agriculture. As a measure of the terms of trade of agriculture, this became an important symbol. "A hundred per cent of parity" became the avowed goal of agricultural policy, even though there is very little reason to suppose that the terms of trade of a given historic period, in this case the period 1909–14, have any ultimate validity as an ideal. Because of differing rates of change in productivity in different parts of the economy, we should expect the terms of trade of different sectors to change. If, for instance, productivity in agriculture rises faster than in the rest of the economy, as it has done in the last thirty years, we would expect the terms of trade of agriculture to "worsen" without any worsening of the incomes of farmers, and without any sense of social injustice.

Even though economic measurement may be abused, its effect on the formation of moral judgments is great, and on the whole I believe beneficial. The whole idea of cost-benefit analysis, for instance, in terms of monetary units, say "real" dollars of constant purchasing power, is of enormous importance in the evaluation of social choices and even of social institutions. We can grant, of course, that the "real" dollar, which is oddly enough a strictly imaginary one, is a dangerously imperfect measure of the quality of human life and human values. Nevertheless, it is a useful first approximation, and in these matters of evaluation of difficult choices it is extremely useful to have some first approximation that we can then modify. Without this, indeed, all evaluation is random selection by wild hunches. It is true, of course, that cost-benefit analysis of all sorts of things, whether of water projects, other pork barrel items, or in more recent years weapon systems, can be manipulated to meet the previous prejudices of people who are trying to influence the decisions. Nevertheless, the fundamental principle that we should count

all costs, whether easily countable or not, and evaluate all rewards, however hard they are to evaluate, is one which emerges squarely out of economics and which is at least a preliminary guideline in the formation of the moral judgment, in what might be called the "economic ethic."

Nevertheless, the economic ethic, or the total cost-benefit principle, is subject to sharp challenge. Two principal criticisms have been made of it, the first of which I think is probably not valid, and the second of which may be valid under limited circumstances. The criticism that I think is not valid is that cost-benefit analyses in particular, or economic principles in general, imply selfish motivation and an insensitivity to the larger issues of malevolence, benevolence, the sense of community and so on. It is quite true, as shown above, that economists have neglected the problem of malevolence and benevolence. Nevertheless, our attitudes towards others can be measured at least as well as we can measure other preferences, either by some principle of "revealed preference" or by direct questioning. It is entirely within the competence of economics, for instance, to develop a concept of the "rate of benevolence" which is the quantity of exchangeables, as measured in real dollars, which a person would be willing to sacrifice in order to contemplate an increase of one real dollar in the welfare of another person. If the rate of benevolence is zero, of course, we have indifference or pure selfishness; if the rate of benevolence is negative we have malevolence, in which case people need compensation in order to contemplate without loss the increased welfare of an enemy, or in reverse would be willing to damage themselves in order to damage another. The rate of malevolence then would be the amount in real dollars one would be prepared to damage one's self in order to damage another person to the extent of one dollar. These rates of malevolence incidentally are frequently quite high. It apparently costs the United States about four dollars to do one dollar's worth of damage in Vietnam, in which case our rate of benevolence towards North Vietnam is at least minus four. In determining cost-benefit analysis we can easily include rates of benevolence and malevolence, adding the benefits and subtracting the costs to those toward whom we are benevolent, multiplied of course by the rate of benevolence, and subtracting the benefits and adding the costs, similarly modifed, to those towards whom we are malevolent.

The concept of a rate of benevolence, incidentally, is at least a partial solution to the perplexing question of interpersonal comparisons of utility around which economists have been doing a ritual dance for at least three generations. Any decision involving other people obviously involves these interpersonal comparisons. They are made, of course, inside the mind of the decision-maker and what his rates of benevolence or malevolence are likely to be is determined by the whole social process in which he is embedded. Surely something can be said about this. We are, for instance, likely to be more benevolent to people who are going to vote for us and perhaps malevolent to people who are going to vote against us. The economic theory of democracy indeed as developed by Anthony Downs and others is a very good example of what I have sometimes called "economics imperialism," which is an attempt on the part of economics to take over all the other social sciences.

The second attack on the "economic ethic" is more fundamental and harder to repulse. This is the attack from the side of what I have elsewhere called the "heroic ethic" [1]. In facing decisions, especially those which involve other people, as

virtually all decisions do, we are faced with two very different frameworks of judgment. The first of these is the economic ethic of total cost-benefit analysis. It is an ethic of being sensible, rational, whatever we want to call it. It is an ethic of calculation. We cannot indeed count the cost without counting. Hence, it is an ethic which depends on the development of measurement and numbers, even if these are ordinal numbers. This type of decision-making, however, does not exhaust the immense complexities of the human organism, and we have to recognize that there is in the world another type of decision-making, in which the decision-maker elects something, not because of the effects that it will have, but because of what he "is," that is, how he perceives his own identity.

This "heroic" ethic takes three major forms—the military, the religious, and the sporting. The heroic ethic "theirs not to reason why, theirs but to do and die" is so fundamental to the operation of the military that attempts to apply an economic ethic to it in the form of cost-benefit analysis or programmed budgeting, or even strategic science as practiced by Herman Kahn, T. C. Schelling, or even Robert McNamara, are deeply threatening to the morale and the legitimacy of the whole military system. Religion, likewise, is an essentially heroic enterprise, even though there is a strong streak of spiritual cost-benefit analysis in it. The enormous role which religion has played in the history of mankind, for good or ill, is based on the appeal which it has to the sense of identity and the sense of the heroic even in ordinary people. "Here I stand and I can do no other" said Luther; "To give and not to count the cost, to labor and ask for no reward" is the prayer of St. Francis. "Do your own thing" is the motto of our new secular Franciscans, the Hippies. In our national religion, President Kennedy said, "Ask not what your country can do for you, ask only what you can do for your country." We find the same principle in poetry, in art, in architecture, which are constantly striving to disengage themselves from the chilling embrace of cost-benefit analysis. I cannot resist quoting here in full what has always seemed to me one of the finest expressions in English poetry of the heroic critique of economics—Wordsworth's extraordinary sonnet on King's College Chapel, Cambridge (Ecclesiastical Sonnet, Number XLIII):

INSIDE OF KING'S COLLEGE CHAPEL, CAMBRIDGE

Tax not the royal Saint with vain expense,
With ill-matched aims the Architect who planned—
Albeit labouring for a scanty band
Of white-robed Scholars only—this immense
And glorious Work of fine intelligence!
Give all thou canst; high Heaven rejects the lore
Of nicely-calculated less or more;
So deemed the man who fashioned for the sense
These lofty pillars, spread that branching roof
Self-poised, and scooped into ten thousand cells,
Where light and shade repose, where music dwells
Lingering—and wandering on as loth to die;
Like thoughts whose very sweetness yieldeth proof
That they were born for immortality.

Okay, boys, bring out your cost-benefit analysis now! There is a story, for the truth of which I will not vouch, that Keynes once asked the chaplain of King's College if he could borrow the chapel for a few days. The chaplain was overjoyed at this evidence of conversion of a noted infidel until it turned out that Keynes had got stuck with a load of wheat in the course of his speculations in futures con-

tracts and wanted to use the chapel for storage.

The "lore of nicely-calculated less or more," of course, is economics. I used to think that high heaven rejected this because its resources were infinite and therefore did not need to be economized. I have since come to regard this view as theologically unsound for reasons which I cannot go into here, but also for a more fundamental reason. High Heaven, at least as it exists and propagates itself in the minds of men, is nothing if not heroic. The power of religion in human history has arisen more than anything from its capacity to give identity to its practitioners and to inspire them with behavior which arises out of this perceived identity. In extreme form, this gives rise to the saints and martyrs of all faiths, religious or secular, but it also gives rise to a great deal of quiet heroism, for instance, in jobs, in marriage, in child rearing and in the humdrum tasks of daily life, without which a good deal of the economy might well fall apart.

A good deal of the criticism of economics from both left and right arises from dissatisfaction with its implied neglect of the heroic. There is a widespread feeling that trade is somehow dirty, and that merchants are somewhat undesirable characters, and that especially the labor market is utterly despicable as constituting the application of the principle of prostitution to virtually all areas of human life. This sentiment is not something which economists can neglect. We have assumed all too easily in economics that because something paid off it was therefore automatically legitimate. Unfortunately, the dynamics of legitimacy are more complex than this. Frequently it is negative payoffs, that is, sacrifices, rather than positive payoffs, which establish legitimacy. It has been the precise weakness of the institutions that we think primarily of as economic, that is, associated with exchange, such as the stock market, the banking system, organized commodity markets and so on, as Schumpeter pointed out, that they easily lose their legitimacy if they are not supported by other elements and institutions in the society which can sustain them as integral parts of a larger community. On the right also we find nationalists, fascists, and the military, attacking the economic man and economic motivation from the point of view of the heroic ethic. It is a wonder indeed that economic institutions can survive at all, when economic man is so universally unpopular. No one in his senses would want his daughter to marry an economic man, one who counted every cost and asked for every reward, was never afflicted with mad generosity or uncalculating love, and who never acted out of a sense of inner identity and indeed had no inner identity even if he was occasionally affected by carefully calculated considerations of benevolence or malevolence. The attack on economics is an attack on calculatedness and the very fact that we think of the calculating as cold suggests how exposed economists are to romantic and heroic criticism.

My personal view is that, especially at his present stage or development, man requires both heroic and economic elements in his institutions, in his learning processes and in his decision-making and the problem of maintaining them in proper balance and tension is one of the major problems of maturation, both of the individual person and of societies. Economic man is a clod, heroic man is a fool, but somewhere between the clod and the fool, human man, if the expression may be pardoned, steers his tottering way.

Let me conclude by stealing another idea from economics and applying it to general moral science. This is the concept

of a production function, some sort of limited relationship between inputs and outputs as expressed in the great biblical principle that grapes are not gathered from thorns, or figs from thistles (Matthew 7:16). There are production functions not only for grapes and figs, but also for goods and bads, and indeed for the ultimate Good. We dispute about what is good, about what outputs we want as a result of the inputs we put in. We dispute also however about the nature of the production functions themselves, what inputs in fact will produce what outputs. In the case of physical production functions the problems can be resolved fairly easily by experimenting, even though there are some pretty doubtful cases, as in the case of cloud seedings, which do not seem to be demonstrably more effective than rain dances. In the case of moral production functions, however, the functions themselves are much in dispute, and there may indeed be more disputation about the production functions than there is about the nature of the desired outputs themselves. I was impressed some years ago, when engaged in a long arduous seminar with some young Russians and young Americans with how easy it was to agree on ultimate goals, even across these widely divergent ideologies, and how extraordinarily hard it was to agree about the inputs which are likely to produce these ultimate goals.

There is a problem here in human learning of how do we get to know the moral production functions in the complex melee of social, political, and economic life, when it seems to be pervaded throughout with a note of almost cosmic irony in which almost everything we do turns out different from what we expect because of our ignorance, so that both the bad and the good we do is all too often unintentional. I cannot solve this epistemological problem in one short paper, but I recommend it as a major intellectual challenge to the moral sciences. What I am concerned with here is with economics as an input into this moral production function. Does economics, as George Stigler has suggested, make people conservative [3]? If so, it is perhaps because it simply points out the difficulties and dangers of heroic action and makes people appreciate the productivity of the commonplace, of exchange and finance, of bankers and businessmen, even of the middle class which our heroic young so earnestly despise. Perhaps this is why so many young radicals today have abandoned economics as a poisoned apple of rationality which corrupts the pure and heroic man of their identities and sympathies. Economics is a reconciler, it brings together the ideologies of East and West, it points up the many common problems which they have, it is corrosive of ideologies and disputes that are not worth their costs. Even as it acts as a reconciler, however, does it not undermine that heroic demand for social mutation which will not be stilled in the voices of our young radicals?

I confess I have been deeply disturbed when I have asked myself these questions and I have no easy answers to them. Nevertheless, I am not sorry that I became an economist, for to belong to a body of people who have never even thought of introducing malevolence into their social theory is somehow in this day and age a little cheering. The anxieties, the moral anguish, and the intense dispute which has racked the American Economic Association this year and which is symbolized by the question as to whether we should move our meeting from Chicago is symptomatic of the fact that not even the study of economics can turn people into purely economic men. Strangely enough it was the mathematical economists and eco-

Implications of Property Rights for Government Investment Choices

By Louis De Alessi*

"To govern, as wise men have said, is to choose."

John F. Kennedy

Economists and others concerned with efficiency in government have been cheered by the increasing use of benefit-cost analysis as a device for discriminating among alternative governmental investment options. That the analysis has some limitations is well known, and there has been considerable interest in clarifying its range of application and improving its usefulness.[1] Nevertheless, it appears that the bias inherent in decisions made in an environment marked by the absence of private property rights has not yet been fully appreciated.[2] This essay, therefore, will focus upon some ambiguities of benefit-cost analysis and upon some implications that the existing structure of opportunities and incentives within the government has for investment choices.

Section I suggests that, in a wider variety of cases than may be generally recognized, benefit-cost analysis cannot yield an unambiguous ranking of the projects under consideration. Moreover, present-value benefit-cost data frequently mask information relevant to the choice. Section II briefly states the utility maximization model and shows some of its implications for decision making within the government. Section III considers some aspects of the current debate about what discount rate should be used in evaluating government investments. Section IV presents a few conclusions.

I

Benefit-cost analysis undoubtedly is capable of generating at least some of the information relevant to the ranking of government investment options by decision makers within the government. Nevertheless, it seems incapable of ranking them unambiguously. If the cost and benefit streams cannot be measured in the market, not even indirectly, it seems clear that no objective[3] basis for discrimination is available. It follows by definition that a unique ordering of the investment options being compared is impossible without reference to the decision maker's utility function. As will be suggested in this section, however, benefit-cost analysis is also incapable of yielding a unique ordering irrespective of whether the choice relates to nonmarketable investments whose cost streams can be evaluated in the market or to marketable investments whose cost and benefit streams can be so evaluated.

* The author is professor of economics at George Washington University. This essay, prepared while the author held a visiting appointment at the Institute for Defense Analyses, benefited from comments by V. L. Broussalian, J. S. McGee, D. G. Davies, W. A. Niskanen, G. Tullock, and an anonymous referee of this *Review*. The views expressed in this paper should not be interpreted as reflecting the official opinion either of IDA or of any of its research sponsors.

[1] One of the most frequently cited summaries of the recent literature is by Prest and Turvey [31]. McKean's thoughtful discussion of benefit-cost analysis and its limitations is still an excellent reference [27].

[2] There are some exceptions, of course. For example, see Broussalian [9], Maas [23], and Lind [22].

[3] Objective means independent of the decision maker responsible for the choice.

Government Investments Yielding Nonmarketable Outputs

The inputs of government investments usually are traded in the market. In such cases, it should be within the state of the art to identify the inputs and to generate a reasonably accurate estimate[4] of the distribution of the probable cost streams associated with each project.[5] Moreover, these cost estimates could be independent of the particular decision maker involved.[6] The streams of benefits, however, are usually much harder to evaluate objectively. In particular, if the outputs are nonmarketable[7] and thus, by implication, cannot be stored, the consumption rate of the benefit stream associated with each investment cannot be altered through time except by changing the production rate itself. The choice of one output stream implies consumption of that and only that stream. Even if nonmarketable outputs can be identified in some meaningful sense (e.g., the capability to deliver a certain number of missiles on target), their valuation is necessarily arbitrary because the choice among the projects in question is essentially a choice among consumption options rather than among investment options, and the value placed upon the differently shaped time streams of benefits will vary among choosers as well as among evaluators.[8] In particular, the explicit benefits (e.g., defense against nuclear attack) usually accrue at different points in time to different groups of individuals outside the organization, and the valuation of these benefit streams is arbitrary. It follows that present-value measures of these streams are similarly arbitrary, and mask useful information. The time profile of the benefit streams matters,[9] and the utility function of the decision maker is relevant to the choice of investment.

If the investments yield nonmarketable outputs, then the time shape of the cost streams, whether measured objectively or not, is also relevant to the choice of investment. Thus, consider two mutually exclusive investment options with identical streams of nonmarketable benefits but different cost streams, where the latter are completely measured in the market. Since the cost streams measure the time pattern of resource withdrawal associated with each investment, different individuals usually bear the cost at different points in time. Unlike the usual investment decisions taken within a private-property market system, the choice of one cost stream over another necessarily involves intertemporal as well as interpersonal utility comparisons, and the decision maker's utility function is relevant to the choice. The

[4] The term "reasonably accurate" is not as vague as may appear. The degree of accuracy to be sought is an economic decision; at some additional cost, more precise estimates may be obtained.

[5] The cost of an investment is the highest-valued alternative foregone as a result of acquiring, maintaining possession of, and operating the investment. But the alternative foregone by whom? The owners? The decision makers? The opportunity cost may vary depending upon which group is relevant, and the latter, in turn, depends upon the system of property rights and the purpose for which the cost information is sought. For purposes of benefit-cost analysis, costs are implicitly defined with reference to the owners of the organization, and are usually measured by the market value of the resources used.

[6] This statement does not deny that the decision maker may derive different streams of private pecuniary and nonpecuniary costs and benefits from the alternative government cost streams he is comparing.

[7] The costs of enforcing and exchanging private property rights in the outputs are sufficiently high to preclude production of the output in the private sector.

[8] For a thorough discussion of the difficulties introduced by nonmarketable investments, see Broussalian [9].

[9] For example, disregarding costs (i.e., suppose that the cost streams are identical), the preference among alternative time patterns of missile capability would depend upon such things as the likely enemy threat at different points in time and the phasing into the defense structure of complementary and substitutable weapon systems. Since the choice of one system is the choice to consume the time pattern of benefits it yields, it would not be independent of the time shape of the benefit stream.

time shape of the cost streams matters, and discounting would mask pertinent information. For prescriptive purposes, therefore, it is not clear why the project with the least present-value cost, where the streams are discounted at some objective rate, should be preferred.[10] This statement does not deny that present-value estimates may be helpful to a decision maker; it does deny that they summarize adequately all the information provided by the size and time shape of the corresponding streams.[11]

As the preceding comments suggest, in general the use of benefit-cost data to justify the ranking of nonmarketable investments may best be viewed as an *ex post* rationalization of the ranking yielded by the decision maker's preference function as constrained by his opportunity set. The time patterns of the cost and benefit streams themselves are relevant to the choice, and the subjective preference of the decision maker for these streams is either masked (if the data are adjusted to reflect his preference) or neglected (if they are not).

[10] A higher-level decision maker has the incentive to impose on choices he delegates to lower-level decision makers the discount rate defined by the slope of his intertemporal budget constraint. The highest-level decision maker within the government bureaucracy (e.g., the President) has the incentive to seek, on investment choices he delegates, the cost stream associated with the least onerous time pattern of taxation for any given stream of benefits. Choices according to this objective would release funds that he could use to buy more of all benefits or decrease taxes or both. Since the pattern of taxation may be modified by either lending or borrowing, a case can be made for discounting such cost streams (measured in the market) at the rate reflecting the cost of borrowing to the government. This approach would be applicable to a wide variety of choices within the government and particularly within the Department of Defense.

[11] The decision maker does not have the option, given the basket of benefits, to choose the project with the least present-value cost and to appropriate the resources saved for his own private consumption or investment program. It follows that only by chance would he be indifferent among projects with identical streams of benefits and the same present-value costs to taxpayers discounted at some objective interest rate.

Government Investments Yielding Marketable Outputs

In some instances, market valuation of both inputs and outputs is feasible, at least in principle.[12] Ranking investments so as to maximize the present value of the net income stream, however, in general is still not a sufficient criterion for establishing the relative desirability of the investment options being compared. Whose wealth is being maximized? If the individuals receiving the benefits and the individuals bearing the costs vary from investment to investment, *any* ranking of the options under consideration involves some interpersonal utility comparison. Ranking projects so as to maximize the present value of the net income stream implies that the redistribution of benefits within the community either is irrelevant or will be done subsequently. The latter course of action is not feasible, since it requires knowledge of individual utility functions. The former, a value judgment, implies that few, if any, marketable outputs would be produced (whether directly or indirectly) by the government. To say that an output is marketable is to say that the costs of transferring property rights in the output and of excluding nonpayers are low enough for production in the private sector to be economical. Thus, if marketable outputs are produced by the government, political and other considerations subjective to the decision maker are inherent in the choice of marketable investments.[13] Benefit-cost analysis is still

[12] The following comments would also be applicable to the marketable portion of mixed investments and, with obvious modifications, to nonmarketable investments.

[13] This point has been recognized for some time, and has led to the suggestion that the expected redistributional effects of each alternative investment be estimated and presented to the decision maker in conjunction with other benefit-cost data. E.g., Krutilla and Eckstein [21], McKean [27], and Haveman [19].

incapable of ranking investment alternatives objectively.

II

The preceding section suggests that, in a wide variety of cases, benefit-cost analysis cannot yield an unambiguous ranking of the investments under consideration. If the pattern of taste of the decision maker is relevant to the choice, the question of bias in these choices may usefully be explored.

The government of any society may best be viewed as composed of many competing individuals with conflicting goals.[14] Individual utility functions presumably reflect a variety of goals such as wealth, power, and knowledge as well as goals such as a taste for the welfare of others. Social welfare, as conceived by the individual, may thus be hypothesized to enter his utility function. For the purposes of this essay, utility functions are hypothesized to contain the same arguments whether the individual is a government official or not. This assertion is neutral with respect to the hypothesis that the relative weights assigned by government officials to the arguments in their utility functions are systematically different from those assigned by the rest of the population. Finally, the individual may be taken to be selfish in the sense that he values the right to make decisions regarding the allocation of the resources which he controls.[15] Even if all government decision makers were motivated by the same lofty ideals, few would place the same relative weights on different goals or on different means of achieving them, and each would have the incentive to further those he prefers. Group decisions may thus be considered the outcome of decisions made by individuals in pursuit of their own self interest.

By definition, a government decision maker, (d), must have some discretion in making choices. The amount of discretion which he has over the allocation of resources under his supervision, by defining the share of public property rights which he enjoys by virtue of his office, is a determinant of his transformation function between pecuniary and nonpecuniary sources of income, and, therefore, of his opportunity set. The weaker the correlation between d's welfare (as *he* sees it) and that of higher level decision makers (as *they* see it) the greater is the extent to which d is likely to increase his own welfare at the expense of the latter. Higher level decision makers (ultimately, the owners—the voters, in the case of democratic government) recognize the existence of this incentive, and impose special constraints upon d's choices. These constraints appear either explicitly or implicitly in the employment contract, and may include such matters as detailed specification of the purposes for which the budget can be spent (e.g., line appropriations).[16] Nevertheless, since detecting and policing infractions are costly, and since higher level decision makers are also utility maximizers, each decision maker has the opportunity to divert some resources to the max-

[14] One of the earliest statements of this general approach was presented by Downs [15], whose current work provides additional insights into bureaucratic behavior [16]. More directly, however, the argument developed in this section represents an extension to government bureaus of recent developments in the theory of the firm [1] [11], and owes much to Alchian and Kessel [6] and to Becker [8].

[15] Since the right to make decisions regarding the allocation of resources affects the welfare of the chooser and is scarce, it is an economic good according to the standard definition of the term. As in the case of any economic good, the lower the cost the more will be demanded.

[16] Exchanges do occur within the public sector [28]. Given the structure of search and transaction costs imposed by existing institutional constraints, however, such exchanges yield a much smaller degree of price uniformity than would prevail within the private sector. By and large, different decision makers within the government will look at different structures of relative prices, including interest rates.

imization of his own utility, sacrificing some of the utility of his hierarchical superiors in order to increase his own by some increment.[17]

The preceding comments are applicable to a bureaucrat within a business firm[18] as well as to a bureaucrat within a government. As Alchian notes [2], the crucial difference between private and public organizations lies in the relatively higher cost of transferring shares of ownership in the latter.[19] The only way an individual can change his portfolio of ownership in government organizations is to move from one jurisdiction to another or to change the structure of the organizations. This procedure is far more expensive than changing one's portfolio of ownership in private organizations, and the opportunity set is more limited—for example, there are few communities in which an individual may own shares in only one or two specific government activities.[20] For the purpose of this essay, therefore, property rights in public organizations may be taken to be nontransferable. The nontransferability of property rights in public organizations rules out individual specialization in their ownership. Whether such specialization would have occurred as some individuals sought to make their wealth more dependent upon their own activities, or as some individuals sought to exploit their comparative advantage in control or risk bearing is not crucial. What is crucial is that, as a result, the owners' incentive to detect and inhibit managerial behavior which they consider undesirable is weaker relative to the incentive existing in privately owned firms.[21]

[17] Even if the goals of decision makers at different hierarchical levels were identical, as the size of the organization increases the quality of the information received and of the instructions provided by higher-level decision makers deteriorate (i.e., the costs of maintaining the same level of quality increase), implying an increased divergence between the decisions taken at lower hierarchical levels and those sought at higher levels [33] [37] [29]—and the rise of "organization men."

[18] Alternative forms of business organization represent, within broad limits, alternative contractual agreements among individual owners and managers as to the sharing of property rights to the resources of the firm. Given the market environment, it may be hypothesized that the owners will choose the wealth-maximizing form of business organization. To the extent that some organizations imply a greater opportunity for utility-maximizing behavior on the part of the managers, the wealth of the owners of the firm will be smaller than it would have been in the absence of such behavior, but greater than it would have been with some other form of organization. Economic theory implies that this must be so, since individuals have an opportunity to choose which type of firm to establish and to own. Although the survival argument [4] is subject to some limitations, it does provide intuitive support for this assertion.

[19] For further discussion of the economics of property rights see, *inter alia,* [5, Ch. 9, 24] [3] [10] [12] [13] [14].

[20] An individual presumably would adjust to any given portfolio of ownership in public organizations by modifying his portfolio of ownership in competing and substitutable private assets.

[21] The consequences of any decision may include both pecuniary and nonpecuniary costs and benefits to (1) the individual who makes the decision, (2) other individuals (e.g., owners) within the same organization, and (3) all other individuals in the community. Insofar as the decision maker is concerned, the only pecuniary and nonpecuniary consequences that are relevant to his decision are those that affect his welfare, including his taste for the welfare of others. All other consequences, to him, are externalities. A system of private property rights, by tying the wealth of the decision maker more closely to the consequences of his decisions, provides an incentive to internalize externalities and to convert nonpecuniary into pecuniary costs and benefits [14].

In general, the behavior of owner-managed, nonregulated competitive business firms may be predicted with a tolerable degree of accuracy even if the nonwealth components of the manager's utility function are disregarded. Application of the wealth-maximization model to a wider class of business firms, however, yields some implications which are falsified by the evidence (e.g., the findings with respect to sales and growth "maximization"). For example, owner-managed monopolies subject implicitly or explicitly to government regulation find that auxiliary constraints upon their profits lower the opportunity cost of nonpecuniary sources of income. Moreover, alternative ownership structures of the firm (e.g., corporations) which weaken the relationship between the wealth of the managers and the wealth of the owners affect managerial choices systematically and predictably, and the wealth maximization hypothesis is no longer adequate.

In turn, this implies that higher level decision makers in public organizations have relatively greater opportunity to pursue, and to translate through subordinates, policies conducive to increasing their own welfare relative to the welfare of the owners as the latter see it. It may be noted that the costs to the owners of enforcing contracts may be expected to vary systematically depending upon the governmental organization in question (e.g., local, federal, independent governmental agency).

The utility maximization hypothesis implies that the individual decision maker (*d*) will seek to expand the resources under his supervision beyond the point which, if the costs of enforcing contracts were zero, would be optimal from the point of vew of higher level decision makers. The higher the cost to superiors of enforcing contracts with a lower level decision maker, other things being the same, the larger the opportunity set of the latter. By increasing the resources under his supervision, a bureaucrat will generate an increase in his marginal product, and, usually, in his salary [25] [26] [36]. Similarly, *d* will have the opportunity to hire more congenial people and to undertake projects he prefers at the cost of some efficiency as viewed by higher level decision makers. Since the streams of pecuniary and nonpecuniary benefits which *d* obtains from his position are coterminous with his stay in office, the expected length of his tenure affects their present value (to him) and, therefore, the exchange ratios among alternatives in his field of choice. All other things being the same, *d*'s choices will reveal a higher time preference if he shares public property rights to some resources than if he owns private property rights to them: the opportunity cost of alternatives foregone beyond his withdrawal from public office is lower. It follows that *d* will have an incentive to shift expenditures from the future toward the present, including earlier investments. As suggested in the next paragraph, this implies that *d* will have an incentive to use as low a discount rate in evaluating investments in his "firm" as his superiors will allow. Moreover, since *d*'s performance is not evaluated by the profitability of his firm, he is induced to offer other evidence of his worth to the organization. For example, the utility maximization hypothesis implies that presidents of tax-supported universities usually allocate relatively more of their resources to building and equipment and relatively less to faculty than presidents of private institutions (buildings and equipment are more readily observed and understood by legislatures and, ultimately, by taxpayers). Similarly, decision makers within the government seek more visible tokens of their accomplishments. The managers of a government unit responsible for building dams will build more and bigger dams earlier in time than the owners would allow if the costs of enforcing contracts were zero.

The utility maximization hypothesis implies, *inter alia,* that government decision makers have the incentive to increase the current overall expenditures of their firms beyond the "optimum" point [30], fall-

---

Attempts to extend the wealth maximization hypothesis to choices within the government yield even poorer predictions; the relationship between the wealth of the managers and the wealth of the voters is so weakened as to be useless. It may be argued that the wealth of the owners is irrelevant in this context, and that the overt objective of government decision makers is to maximize the present value of some benefit or set of benefits for some given cost (or of minimizing such costs for a given basket of benefits), where the benefits to be sought are at least partially given to the organization—proximately, perhaps by some long dead politician, ultimately, at least in democratic societies, by some of the voters, who also evaluate the performance of the managers. That, however, is precisely the point. Public ownership implies that the performance of the managers is even costlier to evaluate and to police, implying an even greater opportunity for utility maximizing behavior on the part of the managers.

ing short or exceeding it for specific projects. Given currently accepted criteria for justifying the ranking of investments, this implies that *d* has the incentive to discount future costs and benefits at as low a discount rate as higher level decision makers will allow.[22] Taking costs first, the lower the discount rate the higher is the present value of costs to be incurred in the future relative to costs incurred in the present, and the lower is the cost of capital goods in the present relative to the present value of operating and other costs in the future.[23] Thus, *ceteris paribus,* lower discount rates favor more capital intensive projects with a higher proportion of costs incurred in the present. Turning to benefits, more capital intensive projects usually have a longer life; since benefits accrue over a longer period, *ceteris paribus,* the lower the discount rate the higher is the present value of the benefit streams, biasing choices in favor of longer-lived, more capital intensive investments. Thus, lower discount rates allow *d* to justify not only a reallocation of the investment budget from the future to the present, but an even bigger present budget.[24]

The implication that the interest rate used to discount government investments is lower than the discount rate that would have been used by a private firm considering a comparable investment does not seem inconsistent with the evidence. Insofar as explicit rates are concerned, the Department of Defense has formally adopted a zero discount rate on a number of occasions, while the Bureau of Reclamation and the Corps of Engineers respectively have used 2.5 and 2.5-3 per cent for years [17, p. 94] [19, p. 97]. The list could easily be extended. Although implicit rates are more difficult to obtain, the current debate regarding why government investments are too capital intensive by private standards is at least suggestive.

To justify the choice of projects which he prefers, *d* has the additional incentive to bias their costs downward and their benefits upwards relative to other projects he is considering.[25] There is considerable room for bias not only in the values to be attached to the items of cost or benefit considered, but also with respect to the items included or excluded. Benefits which are not relevant may be included, the life of the project may be made to appear longer than expected, and so on. The opportunity to bias benefit estimates, however, is much greater than the opportunity to bias cost estimates. First, costs are generally incurred in the market, where they

[22] The presumption that voters and higher-level officials attach at least some weight to present value calculations is suggested by the discounting rules adopted or recommended by various governmental agencies such as the Subcommittee on Benefits and Costs [39], the Bureau of the Budget [40], and the Bureau of Reclamation [41]. See [17, esp. pp. 94-95].

[23] Suppose that project A consists of building a dam today and that project B consists of building the same dam in two stages, the first stage today and the second ten years hence. As the table below shows, the lower the discount rate the lower the present value of A's cost stream relative to the present value of B's cost stream—regardless of the size and time shape of the latter.

Alternative Projects	Cost Streams t	$t+10$	Present-Value Costs 0 Per Cent	5 Per Cent	10 Per Cent
A	200	0	200	200	200
B	100	100	200	161	139

The same relationship holds if A represents a project with given initial investment costs and given operating and other costs in the future, and if B represents a substitutable project with lower initial investment costs and higher operating and other costs in the future.

[24] Although the two approaches are not mutually exclusive, the utility maximization hypothesis seems to offer a much more plausible answer to the question of why governments undertake investments which are "too" long lived by private market standards than the answer, based upon altruism, offered by Marglin [24], Sen [32], and others. The point at issue will be considered more carefully in Section III below.

[25] The analysis does not imply that *d* is greedy or selfish in the vulgar sense of the word. The bias in *d*'s choices may well be due to his sincere belief that certain projects are best for his fellow citizens. The point is simply that *d*'s preferences count.

may be measured directly, whereas the value of most benefits (even if marketable) is imputed. Secondly, costs occur earlier in time relative to benefits. Thus, the costs to higher level decision makers and to competing agencies of detecting (i.e., acquiring and processing the relevant information) and policing bias are less for cost than for benefit estimates. The consequences are obvious. Moreover, as noted earlier, benefit-cost data may be expected to be biased toward larger, more impressive projects.

There is some empirical evidence which, although weak and therefore to be considered with extreme caution, does not appear inconsistent with these implications. For example, Freeman's recent study [18] of six projects undertaken by the Bureau of Reclamation suggests that cost estimates were more accurate than benefit estimates. Although estimated costs exceeded realized costs by about one per cent (one difficulty here is the incentive to adjust actual to budgeted costs), all costs (e.g., the opportunity cost of water) were not included, and Freeman concludes that the economic costs of the projects were slightly understated. According to Freeman, benefits were substantially overstated; for example, he estimates that the actual benefits of one project were negative. Moreover, his observation that adjusted benefit-cost ratios and size of project were significantly but inversely correlated is particularly interesting. Freeman's findings are not presented as a test of the hypotheses advanced in this essay; among other things, if the argument presented in Section I above is correct, his measurement of costs and benefits would require modification. In conjunction with the empirical evidence relating to firms' behavior (e.g., [1]), however, they suggest that extension of the utility maximization approach to government decisions is sufficiently fruitful to justify systematic evaluation.

III

The current debate regarding the discount rate applicable to choices within the government has suffered from the failure to distinguish clearly between the rate(s) which a government official is induced to use by the institutions affecting his choices and the rate(s) which he should use. If the purpose is to predict which projects will be undertaken, the analysis must relate the set of objectives sought by the decision maker (his utility function) to the limitations upon his field of choice (his budget constraint, broadly interpreted). If the purpose is to determine which projects should be undertaken according to some other criteria (e.g., Pareto efficiency), then a case must be made for their relevance, and the analysis must show the consequences of their adoption upon the allocation of resources.

The literature so far has focused upon the issue of which projects should be undertaken, subject to the usually implicit constraint that the costs of adopting and enforcing the chosen criteria of efficiency are zero. Hirshleifer, for example, has argued on Paretian grounds that the discount rate should be the same rate that would be applied, in principle, by a business firm evaluating a comparable investment [20].[26] Comparable investments are defined as those having the same proportionate distribution over time of contingent claims to net income for the same probable future states of the world. A number of economists, on the other hand, have suggested that the discount rate used within the public sector should be lower than the rate currently prevailing in the private sector (e.g., [7]). To the extent that the arguments offered by individuals in this group are not elegant statements of

[26] A price-taking, wealth-maximizing firm and, say, a monopolistic, utility-maximizing firm contemplating comparable investments would not necessarily use the same discount rate. Only price-taking firms presumably would provide an acceptable Paretian benchmark.

personal value judgments, they rest largely upon the belief that variations in transaction, information, and insurance costs prevent equivalent time-state claims from selling at the same price (a phenomenon usually ascribed to "market imperfections"), and thus lead to higher rates in the private sector than would otherwise be the case.[27] Arrow, carrying the issue of market imperfections one step further, has suggested that *both* public and private rates should be equal at a lower discount rate than currently prevails in the market [7]. Even if the alternative techniques for achieving this objective were known, their implementation costs cannot be equal to zero. Moreover, such proposals at best would affect only marketable investments; even then, as the discussion in Section I above implies, their implementation would still fail to yield an unambiguous ranking of the projects involved.

Although the distinction may seem artificial for some purposes, the arguments for using a lower discount rate within the government perhaps should be separated from those seeking to explain why governments in fact do undertake investments which are too long-lived by market standards. In particular, Marglin recently offered an explanation based upon the altruism of present toward future generations.[28] Marglin's statement evoked some sharp criticisms,[29] and subsequently Sen [32] attempted to salvage the argument. Sen considers a community of individuals each of whom must choose one of two alternatives, A or B, where the payoff to each individual is a function of the actions of all individuals. Moreover, the preference ordering of each individual satisfies the two following conditions: (1) given the set of actions of others, the individual is better off doing A rather than B, (2) given the choice between everyone doing A and everyone doing B, each individual prefers the latter to the former. If B stands for saving one more unit for the sake of future generations and A for not doing it, Sen's analysis seems to suggest that altruism unambiguously yields the desired implication. If intended to explain the phenomenon at issue, such an argument must be faulted at least on methodological grounds. The conditions specifying the preference ordering of individuals are not initial hypotheses in the deductive system of economic theory. Rather, they are the antecedent conditions which specify the set of circumstances in which Sen's model is applicable. Thus, the validity of the antecedent conditions must be established empirically. In the absence of such empirical evidence, the model has no explanatory value. There is an infinite number of *ad hoc* assertions consistent with any given set of observations, and, without additional evidence, there is no decision rule that allows the investigator to

[27] It is not wholly clear that the evidence justifies this belief. One problem is that alternative time-state claims which seem equivalent to the investigator may in fact not be equivalent to the buyers and sellers of such claims.

[28] Marglin's argument, based upon a model incorporating altruism as a variable in individual utility functions, is that external effects render an atomistic market inadequate for registering the time preference of individuals [24]. Marglin concludes that the interest rate determined in the market need not have any normative significance in planning collective investments.

[29] Tullock [34] objects to a formulation placing altruism toward the present generation in clear opposition to altruism toward future, wealthier generations. Usher [35] argues that altruism alone is not a sufficient condition for the actual rate of investment to differ from the socially optimal rate as defined by Marglin. Lind [22] suggests: (1) if every individual, given market equilibrium, feels that the marginal rate at which he is willing to substitute present for future consumption is correct for every other individual, then his private and social rates of discount are equal and the market rate is of normative significance for public investment decisions; (2) in general, it is not possible to achieve Pareto optimum by setting the interest rate at any particular level; in this sense, there is no social rate of discount. Lind also argues that, from the standpoint of economic welfare, the distribution of total investments among individuals is as important as the level of total investment, and a political solution is required.

discriminate among the alternative hypotheses. The implications of the utility maximization hypothesis derived in this essay, on the other hand, can draw for indirect support upon the empirical evidence relating to recent extensions of the theory of the firm [1] [38] and to the theory of consumer behavior.

Granting that altruism is an argument of individual utility functions, the utility maximization hypothesis implies that a decision maker, in response both to his altruism and to the effect of the altruism of others upon his burget constraint, will allocate some resources to gifts of various types, including gifts to future generations. In addition to altruism, however, some members of the present generation also derive pecuniary and nonpecuniary income today from the selection of programs supposed to benefit future generations. Given the present institutional framework, the effect of such incentives cannot be negligible. For predictive purposes, the relevant set of relative prices (including the cost of the earlier utilization of resources) is that which confronts the individual decision maker. The cost to *d* of undertaking a project is the utility he foregoes on the other projects he could have undertaken instead, and the usual equilibrium conditions will be fulfilled.

Clarifying the consequences of adopting alternative efficiency criteria undoubtedly facilitates the political task of choosing the relevant criteria. But even if everyone in the community agreed upon the final criteria to be used in selecting projects, there would remain the problem of inducing the government officials responsible for the choice to use the criteria in the appropriate way. Solution of this problem rests upon understanding why projects are undertaken, and, therefore, of the consequences of alternative systems of property rights upon the choices of decision makers within the government. That is, the question "Which government investments should be undertaken?" is a part of the larger question "What system of property rights within the government will bring about (at what cost? to whom?) a more desired (by whom?) allocation of resources?" One objective of this paper is to shift attention to the study of this problem.

IV

Extension of the utility maximization hypothesis to choices within the government, given the existing structure of property rights, suggests that a decision maker *d* has the incentive and the opportunity to shift investment outlays from the future toward the present. This implies that *d* has the incentive to discount the cost and benefit streams of government investments at a lower interest rate than would be used by a decision maker within a private firm considering a comparable investment. Moreover, *d* has the incentive to bias benefit-cost data in favor of projects he prefers, including larger, more impressive projects. The analysis here suggests that the opportunity to bias benefit estimates is relatively greater than the opportunity to bias cost estimates, with obvious consequences. These implications do not appear to be inconsistent with observation, and, *inter alia,* offer a more plausible answer to the question of why governments undertake investments that are too long-lived by private standards than the answer, based upon altruism, recently offered in the literature.

In general, benefit-cost analysis seems incapable of ranking government investment options unambiguously. If the pattern of taste of the decision maker is relevant to the choice, the question of bias in these choices may usefully be explored. But even if benefit-cost analysis could rank options unambiguously, the problem would remain of inducing government of-

ficials to apply the relevant criteria correctly. Solution of this problem would still rest upon understanding the consequences of alternative systems of property rights upon the choices of decision makers within the government. Subject to the limitations noted in this essay, as well as others already well known, benefit-cost analysis undoubtedly is useful.

REFERENCES

1. A. A. ALCHIAN, "The Basis of Some Recent Advances in the Theory of Management of the Firm," *Jour. Indus. Econ.*, Nov. 1965, *14*, 30–41.
2. ———, "Some Economics of Property Rights," mimeograph.
3. ———, "Private Property and the Relative Cost of Tenure," in P. Bradley, ed., *The Public Stake in Union Power*, Charlottesville 1959.
4. ———, "Uncertainty, Evolution, and Economic Theory," *Jour. Pol. Econ.*, June 1950, *58*, 211–21.
5. ——— AND W. R. ALLEN, *University Economics*, 2d ed. Belmont 1967.
6. ——— AND R. A. KESSEL, "Competition, Monopoly, and the Pursuit of Money," in *Aspects of Labor Economics*, Nat. Bur. Econ. Research, Princeton 1962.
7. K. J. ARROW, "Discounting and Public Investment Criteria," paper presented at Water Resources Conference, Fort Collins, Colorado, July 1965.
8. G. BECKER, *The Economics of Discrimination*. Chicago 1957.
9. V. L. BROUSSALIAN, *The Evaluation of Non-Marketable Investments*, Center for Naval Analyses, S. E. G. Research Contribution 9. Arlington, Virginia 1966.
10. R. H. COASE, "The Problem of Social Cost," *Jour. Law and Econ.*, Oct. 1960, *3*, 1–44.
11. R. M. CYERT AND J. G. MARCH, *A Behavioral Theory of the Firm*. Englewood Cliffs 1963.
12. H. DEMSETZ, "The Exchange and Enforcement of Property Rights," *Jour. Law and Econ.*, Oct. 1964, *7*, 11–26.
13. ———, "Some Aspects of Property Rights," *Jour. Law and Econ.*, Oct. 1966, *9*, 61–70.
14. ———, "Towards a Theory of Property Rights," *Am. Econ. Rev.*, Proc., May 1967, *57*, 347–59.
15. A. DOWNS, *An Economic Theory of Democracy*. New York 1957.
16. ———, *Inside Bureaucracy*. Boston 1967.
17. O. ECKSTEIN, *Water Resource Development*. Cambridge 1961.
18. A. M. FREEMAN III, "Adjusted Benefit-Cost Ratios for Six Recent Reclamation Projects," *Jour. Farm Econ.*, Nov. 1966, *48*, 1002–12.
19. R. H. HAVEMAN, *Water Resource Investment and the Public Interest*. Nashville 1965.
20. J. HIRSHLEIFER, "Investment Decision Under Uncertainty: Application of the State-Preference Approach," *Quart. Jour. Econ.*, May 1966, *80*, 252–77.
21. J. V. KRUTILLA AND O. ECKSTEIN, *Multiple Purpose River Development*. Baltimore 1958.
22. R. C. LIND, "The Social Rate of Discount and the Optimal Rate of Investment: Further Comment," *Quart. Jour. Econ.*, May 1964, *78*, 336–45.
23. A. MAAS, "Benefit-Cost Analysis: Its Relevance to Public Investment Decisions," *Quart. Jour. Econ.*, May 1966, *80*, 208–26.
24. S. A. MARGLIN, "The Social Rate of Discount and the Optimal Rate of Investment," *Quart. Jour. Econ.*, Feb. 1963, *77*, 95–111.
25. R. MARRIS, *The Economic Theory of 'Managerial' Capitalism*. New York 1964.
26. T. MAYER, "The Distribution of Ability and Earnings," *Rev. Econ. Stat.*, May 1960, *42*, 189–95.
27. R. N. MCKEAN, *Efficiency in Government Through Systems Analysis*. New York 1958.
28. ———, "The Unseen Hand in Government," *Am. Econ. Rev.*, June 1965, *55*, 496–506.
29. R. J. MONSEN, JR. AND A. DOWNS, "A Theory of Large Managerial Firms," *Jour. Pol. Econ.*, June 1965, *73*, 221–36.
30. W. A. NISKANEN, "The Peculiar Econom-

ics of Bureaucracy," *Am. Econ. Rev.*, Proc., May 1968, *58*, 293–305.

31. A. R. Prest and R. Turvey, "Cost-Benefit Analysis: A Survey," *Econ. Jour.*, Dec. 1965, *75*, 683–735.
32. A. K. Sen, "Isolation, Assurance and the Social Rate of Discount," *Quart. Jour. Econ.*, Feb. 1967, *81*, 112–24.
33. G. Tullock, *The Politics of Bureaucracy*. Washington 1965.
34. ———, "The Social Rate of Discount and the Optimal Rate of Investment: Comment," *Quart. Jour. Econ.*, May 1964, *78*, 331–36.
35. D. Usher, "The Social Rate of Discount and the Optimal Rate of Investment: Comment," *Quart. Jour. Econ.*, Nov. 1964, *78*, 641–44.
36. O. E. Williamson, *The Economics of Discretionary Behavior: Managerial Objectives in a Theory of the Firm*. Englewood Cliffs 1964.
37. ———, "Hierarchical Control and Optimum Firm Size," *Jour. Pol. Econ.*, Apr. 1967, *75*, 123–38.
38. ———, "Managerial Discretion and Business Behavior," *Am. Econ. Rev.*, Dec. 1963, *53*, 1032–57.
39. Federal Inter-Agency River Basin Committee, Subcommittee on Benefits and Costs, *Proposed Practices for Economic Analysis of River Basin Projects*. Washington 1950.
40. U. S. Bureau of the Budget, *Circular A-47*. Washington December 31, 1952.
41. U. S. Department of the Interior, Bureau of Reclamation, *Reclamation Manual*, Vol. XIII. Washington 1952.

Rules for Ordering Uncertain Prospects

By JOSEF HADAR AND WILLIAM R. RUSSELL*

In many theoretical as well as practical situations one is frequently confronted with the necessity (or at least desirability) of making a prediction about a decision maker's preference (choice) between given pairs of uncertain alternatives without having any knowledge of the decision maker's utility function. Various approaches to this and related problems have appeared in the literature, the most popular of which involve comparisons of means and variances of the probability distributions in question. Examples of these approaches are the works of Harry Markowitz [3] and James Tobin [6] [7]. It turns out, however, that since expected utility is, in general, a function of *all* the moments of the probability distribution, rules of comparison involving only two moments are valid only for limited classes of utility functions, or for special distributions. It is well known, for instance, that, if the utility function is quadratic, then only the mean and the variance enter into the calculation of expected utility. In that case the decision maker will always prefer the prospect with the lower variance (if he is a risk averter), given that the prospects under consideration have identical means; or prefer the prospect with the higher mean so long as the variances are the same. More generally, H. M. Markowitz [3] and Marcel Richter [5] have shown that, if the utility function is an nth-degree polynomial, then the first n moments have to be considered. But it is clear that, if two distributions differ in all of their first n moments, then in order to determine preference for these distributions, one needs to know the magnitudes as well as the signs of the n derivatives of the utility function. In general, economists are reluctant to impose this many restrictions on the utility function because of the severe loss of generality incurred by such restrictions.

Thus, except for such special cases as the quadratic utility function, the specification of distributions in terms of their moments is not likely to yield strong results, essentially because information about the moments cannot be utilized efficiently for the purpose of ordering uncertain prospects under conditions in which the utility function is unknown. In this paper we propose two rules which are more powerful than the "moment method." Under the stronger of these rules, distributions may be ordered according to preference, given *any* utility function, while under the weaker rule orderability obtains for any utility function which exhibits nonincreasing marginal utility everywhere.

I. *Preliminary Discussion*

Since in this paper we are concerned only with individuals who maximize expected utility, we may briefly go back to the axioms of expected utility maximization in order to show how these axioms are related to the proposed rules. Let u_a and u_b denote the utilities, respectively, of prospects a and b, and assume $u_a < u_b$. (The prospects a and b may be either payoffs offered with certainty, or uncertain prospects; in the latter case the notation u_a and u_b should be interpreted as expected utilities.) If α and β are probabilities, and $\alpha < \beta$, then it follows from the so-called

* The authors are associate professors of economics at Case Western Reserve University and the University of Kentucky, respectively. They have benefitted from discussions with Richard Evans, Jan Kmenta, William Pierskalla, James Ramsey, and Thomas Saving. They are especially grateful to the referee for his helpful suggestions.

independence axiom that

$$(1) \qquad (1-\beta)u_a + \beta u_b > (1-\alpha)u_a + \alpha u_b.$$ [1]

This preference between two prospects with identical payoffs has an intuitively appealing explanation: between any two uncertain prospects such as the above, an individual will always prefer that prospect in which the probability of receiving the preferred payoff is higher. In fact, it may be helpful to suggest here that the preferred prospect in the above example may be thought of as having been constructed from the inferior prospect by a redistribution of probabilities from the lower payoff to the higher payoff.

Consider now a similar situation for n-payoff prospects. The utilities are denoted by u_i, and they are labeled in the order of their magnitudes; i.e., $u_i > u_j$ if and only if $i > j$. The respective probabilities are denoted by α_i and β_i. The following question may then be posed: Given an uncertain prospect with utilities u_i, $i=1, 2, \cdots, n$, and respective probabilities α_i, $i=1, 2, \cdots, n$, is it possible to construct another uncertain prospect with the same utilities, and with probabilities β_i, $i=1, 2, \cdots, n$, such that it will be preferred to the former prospect by any expected-utility-maximizing individual? The clue to how this can be done is, of course, to be found in the above two-payoff example; namely, by a redistribution of probabilities towards the higher payoffs. Thus, if we assume

$$(2) \qquad \begin{cases} \alpha_j > \beta_j, \alpha_k < \beta_k & \text{where } j < k, \\ \alpha_i = \beta_i & \text{for all } i, i \neq j, k, \end{cases}$$

then the new prospect is preferred to the initial one; i.e.,

$$(3) \qquad \sum_{i=1}^{n} \beta_i u_i > \sum_{i=1}^{n} \alpha_i u_i.$$

This result follows directly from (1); if we subtract

$$\sum_{\substack{i=1 \\ i\neq j \\ i\neq k}}^{n} \alpha_i u_i$$

from both sides of the inequality in (3), then we obtain an inequality between two two-payoff prospects similar to that given in (1).

The concentration of probabilities at the higher payoffs which distinguishes the preferred prospects from the inferior ones in both (1) and (3) can be given a more formal characterization. We note that $\beta_i \leq \alpha_i$, $i=1, 2, \cdots j, \cdots, k-1$, by assumption (2), hence

$$\sum_{i=1}^{k-1} \beta_i < \sum_{i=1}^{k-1} \alpha_i.$$

But since we also have $\beta_i = \alpha_i$, $i=k+1, k+2, \cdots, n$, it follows that

$$\sum_{i=1}^{r} \beta_i = \sum_{i=1}^{r} \alpha_i, r = k, k+1, \cdots, n.$$

Therefore the following condition holds:

$$\sum_{i=1}^{r} \beta_i \leq \sum_{i=1}^{r} \alpha_i \qquad r = 1, 2, \cdots, n.$$

In words, the value of the cumulative distribution of the preferred prospect never exceeds that of the inferior prospect. This condition is known as *stochastic dominance*, and so we say that the preferred distributions (prospects) in examples (1) and (3) are *stochastically larger* than the inferior distributions (prospects).

It is clear that, in the absence of any specification of the utility function, the construction of a prospect that is preferred to any given prospect can be accomplished only by making use of condition (1), i.e., by a redistribution of probabilities toward the higher payoffs. While we demonstrated only one such construction in example (3), it should be obvious that by following this procedure iteratively, one can construct as many preferred prospects as one may desire; and it will, of course, be true, be-

[1] For the proof see John von Neumann and Oskar Morgenstern [8, p. 618].

cause of the transitivity of preferences, that all such prospects will be preferred to the original prospect. The above illustrations clearly show that there exists a close relationship between stochastic dominance, on the one hand, and preference between uncertain prospects on the other. In fact, this relationship can be stated more precisely in the form of two propositions, each of which is the converse of the other:

Proposition 1. Given any two prospects P and P', if P is stochastically larger than P', then P is preferred to P', regardless of the specifications of the utility function.

Proposition 2. Given any two prospects P and P', if P is preferred to P' for all utility functions, then P is stochastically larger than P'.

The above propositions have been proved by James Quirk and Rubin Saposnik [4] for prospects with a finite number of payoffs. We provide proofs for discrete as well as continuous distributions. But more important, we show that the set of orderable distributions can be enlarged by placing a (frequently invoked) restriction on the utility function. The second set of results of this paper shows that, if the utility function has nonincreasing marginal utility everywhere, then propositions such as the above can be proved for pairs of distributions which satisfy a condition that is weaker than stochastic dominance.

II. *Formal Results*

First, we introduce the concept of stochastic dominance in a formal manner. Strictly speaking, since we are using two different types of dominance in this paper, we shall modify the accepted terminology slightly. Let f and g denote the probability functions of a (discrete or continuous) random variable taking the values x_i, or x, and let $F(x_i)$ and $G(x_i)$ be the respective cumulative distributions. Throughout the paper we shall adopt the convention of labeling the values of the random variable in accordance with their magnitudes; i.e., $x_i > x_j$ if and only if $i > j$. The set of the x_i is denoted by X.

The first definition is that of *first-degree stochastic dominance* (FSD) which is the same concept as that which in the literature is known simply as stochastic dominance.[2] This condition holds between any two distributions whenever one cumulative distribution lies entirely, or partly, above the other.

Definition 1. The probability function g is said to be at least as large as f in the sense of FSD if and only if

$$G(x_i) \leq F(x_i) \qquad \text{for all } x_i \in X.$$

The second dominance condition is weaker than FSD, and we refer to it as *second-degree stochastic dominance* (SSD). It holds whenever the area under one cumulative distribution is equal to, or larger than, that under the other cumulative distribution.

Definition 2. The probability function g is said to be at least as large as f in the sense of SSD if and only if

$$\sum_{i=1}^{r} G(x_i)\Delta x_i \leq \sum_{i=1}^{r} F(x_i)\Delta x_i \quad \text{for all } r < n,$$

where $\Delta x_i = x_{i+1} - x_i$, and x_n is the largest value taken by the random variable. If the random variable is continuous, taking its value in the closed interval $I = x_1 - x_n$, then the above inequality takes the form

$$\int_{x_1}^{x} G(y)dy \leq \int_{x_1}^{x} F(y)dy \qquad \text{for all } x \in I.$$

It is easy to verify that FSD implies SSD.

Throughout the paper it is assumed that the random variable is one-dimensional.

1. *Orderability under FSD*

In this section we are concerned with preference among uncertain prospects under conditions in which the utility function is not restricted in any essential

[2] For the definition and certain applications of this concept see e.g. [1] and [2].

fashion. More precisely, we shall consider utility functions $u=\phi(x)$ that are members of the set U_1, where $U_1=\{\phi(x):\phi\in D_1(x_1,x_n),\ \phi'>0\}$, $D_1(x_1, x_n)$ represents the set of all functions that are continuous and have a continuous first derivative in the closed interval $I=x_1-x_n$, and $\phi'=d\phi/dx$. At first we confine ourselves to probability distributions in which the random variable assumes only a finite number of distinct values, the latter being denoted by x_i, $i=1, 2, \cdots, n$. We consider any two discrete probability functions defined on X, and write these as $f(x_i)=\alpha_i$, and $g(x_i)=\beta_i$. The respective cumulative distributions are given by

$$F(x_i) = \sum_{r=1}^{i} f(x_r), \quad \text{and}$$

$$G(x_i) = \sum_{r=1}^{i} g(x_r).$$

Theorem 1. If g is at least as large as f in the sense of FSD, then g is at least as preferred as f; i.e., if $G(x_i)\leq F(x_i)$ for all $x_i\in X$, then $\bar{u}_g-\bar{u}_f\geq 0$ for all $\phi\in U_1$, where

$$\bar{u}_f = \sum_{i=1}^{n} \alpha_i\phi(x_i), \quad \text{and}$$

$$\bar{u}_g = \sum_{i=1}^{n} \beta_i\phi(x_i).$$

Proof. Using the Theorem of the Mean, it can be shown that

$$(4)\quad \phi(x_i) = \phi(x_n) - \sum_{r=i}^{n-1} \phi'(\xi_r)\Delta x_r \qquad i = 1, 2, \cdots, n-1,$$

where ξ_r is a properly chosen point between x_r and x_{r+1}, and Δx_r as defined in Definition 2. Since

$$\bar{u}_g - \bar{u}_f = \sum_{i=1}^{n} (\beta_i - \alpha_i)\phi(x_i),$$

we have

$$(5)\quad \bar{u}_g - \bar{u}_f = \sum_{i=1}^{n} \Big[\phi(x_n) - \sum_{r=i}^{n-1} \phi'(\xi_r)\Delta x_r\Big](\beta_i - \alpha_i)$$

$$= - \sum_{r=1}^{n-1} \phi'(\xi_r)[G(x_r) - F(x_r)]\Delta x_r \geq 0.$$

The above inequality follows directly from the fact that $\phi'>0$ and $\Delta x_r>0$ by definition, while $[G(x_r)-F(x_r)]\leq 0$ by the hypothesis of FSD. It is also clear that, given the dominance condition, g will be *strictly* preferred to f if and only if f and g are different functions, in which case we have $[G(x_r)-F(x_r)]<0$ for at least one r.

It should be pointed out here that a result such as (5), or (10) below, may be obtained not only for differences between expected utilities, but also for differences between the expectations of any monotonically increasing function (since monotonicity was the only property of $\phi(x)$ used in the above proof). In particular, since the functions x^k, $k=1, 3, 5, \cdots$, are monotonic functions for all $x\in I$, it follows immediately that, if g is at least as large as f in the sense of FSD, then all the odd moments around zero of g are at least as large as the respective moments of f. In the special case in which the domain of definition I is a subset of the nonnegative half-line, then *all* the moments around zero of the dominant distribution are at least as large as the respective moments of the inferior distribution.

We now prove the converse of Theorem 1.

Theorem 2. If g is preferred to f for all utility functions, then g is larger than f in the sense of FSD; i.e., if $\bar{u}_g-\bar{u}_f>0$ for all $\phi\in U_1$, then $G(x_i)\leq F(x_i)$ for all $x_i\in X$, the strict inequality holding for at least one $x_i\in X$.

Proof. The proof is by contradiction. It is shown that, if the conclusion of the theorem fails to hold, then there exists a utility function in U_1 for which the hypothesis of the theorem is contradicted. For this pur-

$$\phi(x) = \begin{cases} \dfrac{x_k^2(b-a)}{2(x_k - x_1)} + \dfrac{(ax_k - bx_1)x}{(x_k - x_1)} + \dfrac{(b-a)x^2}{2(x_k - x_1)} & \text{for } x_1 \le x \le x_k, \\ bx & \text{for } x_k \le x \le x_n, \end{cases} \tag{6}$$

$$\phi'(x) = \begin{cases} \dfrac{(ax_k - bx_1)}{(x_k - x_1)} + \dfrac{(b-a)x}{(x_k - x_1)} & \text{for } x_1 \le x \le x_k, \\ b & \text{for } x_k \le x \le x_n. \end{cases} \tag{7}$$

$$\bar{u}_g - \bar{u}_f = -\sum_{r=1}^{k-1} (A + B\xi_r)[G(x_r) - F(x_r)]\Delta x_r - \sum_{r=k}^{n-1} b[G(x_r) - F(x_r)]\Delta x_r, \tag{8}$$

where

$$A = \frac{(ax_k - bx_1)}{(x_k - x_1)}, \quad \text{and} \quad B = \frac{(b-a)}{(x_k - x_1)}.$$

---

pose it is sufficient to assume that there exists $x_k \in X$, $x_k > x_1$ such that

$$G(x_i) \ge F(x_i) \qquad \text{for } x_1 \le x_i \le x_k,$$

and

$$G(x_i) \le F(x_i) \qquad \text{for } x_k \le x_i \le x_n.$$

Consider the utility function (6) shown above. The parameters a and b may be chosen freely, subject only to the condition $a > b > 0$. Differentiating (6) with respect to x gives equation (7). It is easily verified that $\phi \in U_1$. But by virtue of (5) we now have equation (8). Clearly, the first sum on the right-hand side of (8) is nonnegative, while the second sum is nonpositive. Furthermore, each of the terms $(A + B\xi_r)$ is an increasing function of the parameter a, hence the magnitude of the first sum can be made arbitrarily large by making a sufficiently large. Therefore there exists a utility function in U_1 for which $\bar{u}_g - \bar{u}_f \le 0$.[3]

Since many economic variables (such as profit, income, rate of return, wealth, etc.) are continuous in nature, we feel that it is useful to provide also the continuous versions of the above theorems. For this purpose we need not redefine the concept of FSD, except that we now think of f and g as probability density functions. The latter, as well as the utility function $\phi(x)$, is defined on the interval $I = x_1 - x_n$.[4]

Theorem 1'. If $G(x) \le F(x)$ for all $x \in I$, then $\bar{u}_g - \bar{u}_f \ge 0$ for all $\phi \in u_1$.

Proof. By definition

$$\bar{u}_g - \bar{u}_f = \int_{x_1}^{x_n} \phi(x)[g(x) - f(x)]dx. \tag{9}$$

Integrating by parts gives

$$\begin{aligned} \bar{u}_g - \bar{u}_f &= \phi(x)[G(x) - F(x)]\Big|_{x_1}^{x_n} \\ &\quad - \int_{x_1}^{x_n} \phi'(x)[G(x) - F(x)]dx \\ &= -\int_{x_1}^{x_n} \phi'(x)[G(x) - F(x)]dx \ge 0. \end{aligned} \tag{10}$$

Here we see again that g will be strictly preferred to f if and only if the two density functions are not identical.

Theorem 2'. If $\bar{u}_g - \bar{u}_f > 0$ for all $\phi \in U_1$, then $G(x) \le F(x)$ for all $x \in I$, the strict in-

[3] For alternative proofs of Theorems 1 and 2 see Quirk and Saposnik [4].

[4] The assumption that the range of the random variable is finite is not really a serious assumption, since the range can always be made as large as one may desire. After all, the maximum amount of income or wealth that an individual may either gain or lose is always finite. As a matter of fact, however, the results can also be proved for conditions under which the domain of definition is allowed to go to infinity. But since the proof then requires the introduction of fairly technical arguments that are likely to becloud the conceptual substance of the proposed rules, we felt that it was desirable not to do so in the present paper.

equality holding for at least one $x \in I$.

Proof. The proof is similar to that of Theorem 2. Suppose that there exists $x_k \in I$, $x_k > x_1$ such that $G(x) \geq F(x)$ for $x_1 \leq x \leq x_k$, and $G(x) \leq F(x)$ for $x_k \leq x \leq x_n$. Then for the utility function given in (6) we have by virtue of (10)

$$\begin{aligned} \bar{u}_g - \bar{u}_f &= -\int_{x_1}^{x_k} (A + Bx)[G(x) - F(x)]dx \\ (11) \qquad & \quad - \int_{x_k}^{x_n} b[G(x) - F(x)]dx, \end{aligned}$$

A and B as defined in the proof of Theorem 2. It is obvious again that, by choosing a sufficiently large, the expression in (11) can be made nonpositive.

The most important implication of Theorems 1 and 2 (or 1′ and 2′) is the fact that, when no restrictions are placed on the utility function, then FSD is necessary and sufficient for preference among pairs of uncertain prospects. In other words, FSD is the weakest condition that will guarantee preference for one prospect over another for all utility functions in the set U_1. The significance of this result is discussed further in Section III below. At this point we turn to the problem of ordering uncertain prospects under the weaker of the two dominance conditions.

2. *Orderability under SSD*

In this section we consider utility functions with nonincreasing marginal utility. More precisely, we are concerned with preference among uncertain prospects for all utility functions $u = \phi(x)$ in the set U_2, where $U_2 = \{\phi(x): \phi \in D_2(x_1, x_n), \phi' > 0, \phi'' \leq 0\}$, $D_2(x_1, x_n)$ is the set of all continuous functions with continuous derivatives of order one and two in the closed interval $I = x_1 - x_n$, and $\phi'' = d^2\phi/dx^2$. Utility functions that are members of U_2 exhibit what is known as weak global risk aversion.[5] We prove theorems similar to those in Section 1 above, starting with the discrete case.

Theorem 3. If g is at least as large as f in the sense of SSD, and marginal utility is everywhere nonincreasing, then g is at least as preferred as f, i.e., if

$$\sum_{i=1}^{r} G(x_i)\Delta x_i \leq \sum_{i=1}^{r} F(x_i)\Delta x_i$$

for all $r < n$, then $\bar{u}_g - \bar{u}_f \geq 0$ for all $\phi \in U_2$.

Proof. Applying the Theorem of the Mean in a fashion similar to that which led to equation (4), we have

$$(12) \qquad \phi'(\xi_r) = \phi'(\xi_{n-1}) - \sum_{s=r}^{n-2} \phi''(\eta_s)\Delta\xi_s \qquad r = 1, 2, \cdots, n-2,$$

where η_s is a properly chosen point between ξ_s and ξ_{s+1}, and $\Delta\xi_s = \xi_{s+1} - \xi_s$. Substituting for the $\phi'(\xi_r)$ in (5) from (12) yields

$$\begin{aligned} \bar{u}_g - \bar{u}_f &= -\sum_{r=1}^{n-1} [\phi'(\xi_{n-1}) \\ (13) \qquad & \quad - \sum_{s=r}^{n-2} \phi''(\eta_s)\Delta\xi_s][G(x_r) - F(x_r)]\Delta x_r \\ &= -\phi'(\xi_{n-1}) \sum_{r=1}^{n-1} [G(x_r) - F(x_r)]\Delta x_r \\ & \quad + \sum_{s=1}^{n-2} \phi''(\eta_s)\Delta\xi_s \sum_{r=1}^{n-1} [G(x_r) - F(x_r)]\Delta x_r \geq 0. \end{aligned}$$

We may point out one implication of the result in (13): if SSD holds between any two distributions, then, just as in the case of FSD, the mean of the dominant distribution is at least as large as that of the other distribution, but the SSD condition imposes no general restrictions on the relative magnitudes of the higher moments of the distributions in question. It follows, therefore, that the set of distributions that can be ordered by means of SSD is, in general, larger than that which may be

[5] That is, a decision maker whose preferences are represented by a utility function in U_2 will never prefer an uncertain prospect P over a certain payoff which is equal to the expected value of P.

$$(14)\qquad \phi(x) = \begin{cases} -\dfrac{ax_k^3}{6(x_k - x_1)} + \left[\dfrac{ax_k^2}{2(x_k - x_1)} + b\right]x - \dfrac{ax_kx^2}{2(x_k - x_1)} + \dfrac{ax^3}{6(x_k - x_1)} & \\ & \text{for } x_1 \leq x \leq x_k, \\ bx & \text{for } x_k \leq x \leq x_n, \end{cases}$$

$$(15)\qquad \phi'(x) = \begin{cases} \dfrac{ax_k^2}{2(x_k - x_1)} + b - \dfrac{ax_kx}{(x_k - x_1)} + \dfrac{ax^2}{2(x_k - x_1)} & \text{for } x_1 \leq x \leq x_k, \\ b & \text{for } x_k \leq x \leq x_n, \end{cases}$$

$$(16)\qquad \phi''(x) = \begin{cases} -\dfrac{ax_k}{(x_k - x_1)} + \dfrac{ax}{(x_k - x_1)} & \text{for } x_1 \leq x \leq x_k, \\ 0 & \text{for } x_k \leq x \leq x_n. \end{cases}$$

ordered by means of FSD. Moreover, SSD is also necessary for preference among uncertain prospects for all utility functions in U_2. Hence we can prove the converse of Theorem 3.

Theorem 4. If g is preferred to f for all utility functions with nonincreasing marginal utility, then g is larger than f in the sense of SSD; i.e., if $\bar{u}_g - \bar{u}_f > 0$ for all $\phi \in U_2$, then

$$\sum_{i=1}^{r} G(x_i)\Delta x_i \leq \sum_{i=1}^{r} F(x_i)\Delta x_i$$

for all $r < n$, the strict inequality holding for at least one value of r.

Proof. We once more use a proof by contradiction. To simplify notation let

$$H(x_i) = \sum_{r=1}^{i} [G(x_r) - F(x_r)]\Delta x_r.$$

Suppose that there exists $x_k \in X$, $x_k > x_1$ such that $H(x_i) \geq 0$ for $x_1 \leq x_i \leq x_k$, and $H(x_i) \leq 0$ for $x_k \leq x_i \leq x_n$. Consider the utility function shown in (14) above, where a and b are freely chosen positive parameters. Differentiating (14) gives (15) and differentiating (15) yields (16) above. It is obvious that $\phi(x) \in U_2$. Now, however, we have

$$(17)\qquad \bar{u}_g - \bar{u}_f = -\phi'(\xi_{n-1})H(x_{n-1}) + \sum_{s=1}^{k-2} (A + B\eta_s)H(x_{s-1}),$$

where

$$A = \frac{-ax_k}{(x_k - x_1)} \quad \text{and} \quad B = \frac{a}{(x_k - x_1)}.$$

The first term on the right-hand side of (17) is nonnegative, while the sum in (17) is nonpositive. But each of the terms $[A + B\eta_s]$ is an increasing function of the parameter a, hence the magnitude of the sum can be made as large as desired by choosing a sufficiently large. Therefore there exists a utility function in U_2 for which the hypothesis of the theorem is contradicted.

The last two theorems are the continuous versions of Theorems 3 and 4.

Theorem 3'. If

$$\int_{x_1}^{x} G(y)dy \leq \int_{x_1}^{x} F(y)dy$$

for all $x \in I$, then $\bar{u}_g - \bar{u}_f \geq 0$ for all $\phi \in U_2$.

Proof. Integration by parts yields

$$(18)\qquad \int_{x_1}^{x_n} \phi'(x)[G(x) - F(x)]dx$$

$$= \phi'(x) \int_{x_1}^{x} [G(y) - F(y)]dy \Big|_{x_1}^{x_n}$$

$$- \int_{x_1}^{x_n} \phi''(x) \int_{x_1}^{x} [G(y) - F(y)]dydx.$$

Substituting for the integral in (10) from (18) gives

$$(19)\quad \bar{u}_g - \bar{u}_f = -\phi(x)'\int_{x_1}^{x}[G(y) - F(y)]dy\Big|_{x_1}^{x_n} + \int_{x_1}^{x_n}\phi''(x)\int_{x_1}^{x}[G(y) - F(y)]dydx \geq 0.$$

Theorem 4'. If $\bar{u}_g - \bar{u}_f > 0$ for all $\phi \in U_2$, then

$$\int_{x_1}^{x} G(y)dy \leq \int_{x_1}^{x} F(y)dy$$

for all $x \in I$, the strict inequality holding for at least one $x \in I$.

Proof. The proof is similar to that of Theorem 4. To simplify notation let

$$H(x) = \int_{x_1}^{x}[G(y) - F(y)]dy.$$

Suppose that there exists $x_k \in I$, $x_k > x_1$ such that $H(x) \geq 0$ for $x_1 \leq x \leq x_k$, and $H(x) \leq 0$ for $x_k \leq x \leq x_n$. Then for the utility function given in (14) we have by virtue of (19)

$$(20)\quad \bar{u}_g - \bar{u}_f = -bH(x_n) + \int_{x_1}^{x_n}(A + Bx)H(x)dx,$$

A and B as defined in the proof of Theorem 4. It is clear that, by choosing a sufficiently large, the expression in (20) can be made nonpositive.

III. *Summary and Conclusion*

One of the objectives of this paper was to suggest a set of rules for ordering uncertain prospects; that is, to specify conditions which will permit us to make predictions about preference. The single most important property of these rules is that they are not only sufficient, but also necessary for the respective class of utility functions. For example, in the absence of any restriction on the utility function (except monotonicity), we find that FSD implies preference. But at the same time, if an uncertain prospect is known to be preferred over another prospect for all monotonic utility functions, then FSD must hold. This means essentially that a state of preference implies something about the characteristics of the prospects in question. Thus by means of the FSD condition we can not only predict preference, but we can also make a statement about the characteristics of the uncertain prospects. For instance, we have indicated in the paper that FSD implies a certain relationship between the odd moments (and sometimes also between the even moments) of the prospects under consideration. Consequently, given that P is preferred to P' for all monotonic utility functions, we can immediately say that all the odd moments around zero of P are larger than the respective moments of P'.

Similar relationships hold with respect to the SSD condition if the utility function is assumed to be concave. In fact, since SSD is a weaker condition than FSD, it is capable of ordering a larger set of distributions than that which is orderable under FSD. Indeed, the theorems that make use of SSD may be more important than those involving FSD because of the central position occupied by the concavity assumption. Not only is this assumption widely used in the literature, it is in fact a necessary condition for the existence of a maximum in a large class of problems involving the maximization of expected utility. Certainly, it is obvious that any result within the framework of the theory of risk aversion can be established directly by means of SSD. Conversely, any case of preference under risk aversion must imply SSD; if the latter condition fails to hold, then the result must necessarily be due to a special assumption about the functional form of the utility function.

Since FSD and SSD are both necessary and sufficient, rules involving the latter conditions must be superior to those involving comparisons of moments. The latter rules, as is well known, yield conclusive results only for a special class of utility functions (e.g., quadratic), or for special distributions (e.g., those depending only on mean and variance). The generality of the latter rules is, therefore, severely restricted. In fact, in those cases in which the application of the moment method is made possible by confining consideration to a special class of distributions, the utility function being either unrestricted or assumed concave, any determinate results must imply SSD, or even FSD, depending on the assumption about the utility function. If so, the same results can be established by using one of the dominance conditions directly, thereby obviating the necessity of imposing a restriction on the class of admissible distributions. Since no such restriction is required in the application of the dominance conditions, it is clear that the moment method is, on the whole, less powerful as a means of ordering uncertain prospects.

The weakness inherent in methods involving comparisons of means and variances also manifests itself in Markowitz's efficiency frontier. For example, suppose we consider the case of global risk aversion. Then the true efficiency frontier consists of a set of prospects with no two prospects in that set satisfying the SSD condition (since otherwise there is preference, in which case both prospects cannot be on the frontier). But the Markowitz frontier may not satisfy this requirement. All we know about the latter frontier is that, between any two prospects on the frontier, one has a higher mean as well as a higher variance than the other prospect. However, it is not difficult to think of (or construct) examples of distributions where the one with the higher mean and variance is larger than the other in the sense of SSD (or FSD). It may, therefore, very well be the case that some pairs of distributions on the Markowitz frontier are in fact orderable by means of either SSD or FSD, in which case at least one prospect of each such pair does not belong on the frontier. Conversely, prospects that are off the Markowitz frontier could possibly be members of the true frontier set inasmuch as they may be neither larger nor smaller in the sense of either SSD or FSD than any other prospect on the frontier.[6]

On the practical level, the use of either FSD or SSD should prove to be the most direct and efficient approach to a large number of specific problems. As was pointed out in Section I, the imposition of FSD is equivalent to a redistribution of probabilities from lower payoffs to higher payoffs; in terms of the cumulative distributions, it amounts to lowering the values of the cumulative distribution (relative to some other distribution) for some values of the random variable without raising them at any other points. Consequently, in order to test for the existence of FSD one needs only examine and compare the cumulative distributions of the prospects under consideration. The SSD condition, on the other hand, places a restriction on the areas under the respective cumulative distributions, and hence it may be applied with equal facility. We thus see that both FSD and SSD are specifications of the type that may be easily applied to actual problems since they are defined in terms of the very same concepts which are customarily used for the description of uncertain prospects.

In concluding, we may emphasize again that the superiority of the FSD and SSD conditions derives directly from the fact that these conditions are both necessary and sufficient, given the class of admissible

[6] For an example see Quirk and Saposnik [4].

utility functions. Thus, in the absence of any specification of the utility function, to say that prospect P is larger than P' in the sense of FSD is *equivalent* to saying that P is preferred to P' for all monotonic utility functions; and given risk aversion, to say that P is larger than P' in the sense of SSD is *equivalent* to saying that P is preferred to P' for all concave utility functions. It is because of this equivalence relationship that the dominance conditions convey information which is more essential to the orderability of uncertain prospects than the information obtained from a comparison of moments. In fact, it is clear that any other rule that may yet be proposed cannot yield results that are stronger than those obtained from the use of either FSD or SSD. Thus any result about preference between uncertain prospects which is inconsistent with the theorems presented in this paper must be due to a more restrictive assumption about the utility function.

REFERENCES

1. S. KARLIN, "Dynamic Inventory Policy with Varying Stochastic Demands," *Manag. Sci.*, April 1960, *6*, 231–58.
2. E. L. LEHMANN, "Ordered Families of Distributions," *Annals Math. Stat.*, Sept. 1955, *26*, 399–419.
3. H. M. MARKOWITZ, *Portfolio Selection*. New York 1959.
4. J. P. QUIRK AND R. SAPOSNIK, "Admissibility and Measurable Utility Functions," *Rev. Econ. Stud.*, 1962, *29*, 140–46.
5. M. K. RICHTER, "Cardinal Utility, Portfolio Selection and Taxation," *Rev. Econ. Stud.*, June 1960, *27*, 152–66.
6. J. TOBIN, "Liquidity Preference as Behavior Toward Risk," *Rev. Econ. Stud.*, Feb. 1958, *25*, 65–86.
7. ———, "The Theory of Portfolio Selection," in F. H. Hahn and F. P. R. Brechling, eds., *The Theory of Interest Rates*, London 1965.
8. J. VON NEUMANN AND O. MORGENSTERN, *Theory of Games and Economic Behavior*. New York 1964.

Bargaining Theory, Trade Unions, and Industrial Strike Activity

By Orley Ashenfelter and George E. Johnson*

The purpose of this paper is to examine certain received theories of the firm, trade union behavior, and bargaining in order to derive testable implications concerning the conditions under which labor disputes are more likely to occur. There are at least three reasons why an investigation of industrial strike activity seems fruitful. First, it might be argued that, because of their relatively frequent disruption of key sectors of the economy, work stoppages are the most important public policy issue raised by the existence of trade unions. It would therefore be useful to know whether, as Hicks thought, "... the majority of actual strikes are doubtless the result of faulty negotiation" [18, p. 146] or whether they are an inevitable part of the functioning of an institutionalized market economy. Although it has long been known that the level of strike activity follows the business cycle [23] [20] [29], this leaves open the questions of the behavioral relations involved and their stability over time. Second, data on industrial disputes provide a potentially rich source of material for testing the implications of bargaining theories which purport to explain the outcome of labor-management negotiations.[1] Yet little work seems to have been done to date on the application of bargaining theoretic models to nonexperimental data. A third reason for undertaking a study of this problem stems from the continuing interest in the effect of unions upon both the relative wage structure and the rate of change of aggregate money wages. Most union "power" is derived from the threat of the strike, and, accordingly, we agree with Charles Holt's recent suggestion that "... the theory and analysis of industrial disputes may help to clarify the role that unions play in the determination of wages" [19, p. 50].

I. *A Theoretical Formulation*

Most bargaining models are addressed to a general two-party situation, e.g., bilateral monopoly, in which conventional economic theory fails to lead to a predictable outcome of the terms on which agreements will be reached. If one views labor-management negotiations from this point of departure it is usually difficult to derive any testable implications concerning the conditions under which the parties will fail to agree on a new contract prior to the point at which the previous contract expires. It is generally assumed that the union attempts to maximize some utility or objective quantity, for example the discounted value of its members' wage income over the length of the contract.[2] Likewise the firm would attempt to maximize some objective, say the discounted value of the

* The authors, lecturer and assistant professor of economics at Princeton University and the University of Michigan, respectively, are indebted to J. Cross, H. Levinson, J. Pencavel, A. Rees, and D. Saks for comments and suggestions. At the time this paper was completed Johnson was a visiting research associate at Princeton participating in the Princeton Systems Analysis of the Labor Market Project, financed through OMPER of the U. S. Department of Labor. Only the authors are responsible for the views expressed in this paper.

[1] The two best-known theories of this sort are those of J. R. Hicks [18] and F. Zeuthen [32]. A very readable summary and critique of several more modern bargaining models is contained in Bishop [5].

[2] For some other possibilities see Dunlop [9].

future profit stream. The desire of a particular party to concede or hold out would then depend upon: (1) a set of objective factors such as the state of product demand, the elasticities of labor demand and capital-labor substitution, etc., as well as (2) a set of subjective factors such as the assessment of the bargaining strategy of the other party and attitudes toward risk-taking.[3] Since the failure to agree on a settlement prior to the point of contract expiration is costly to both parties (loss of the wage bill for the union and current profits for the firm), there should be a tendency for the parties to adjust their positions in such a way that they come to an agreement in time to avert a strike. It is the determination or solution of the terms of this agreement prior to a strike to which most bargaining theories appear to be addressed.

There are two explanations expressed in the bargaining literature regarding the reason why strikes take place at all. First, there is some possibility that one party will misjudge the other's intentions and that a strike will result. This view of strike activity, attributable to Hicks, may be summed up by the statement that "... adequate knowledge will always make a settlement possible" [18, p. 147],[4] although some strike activity is inevitable if trade unions are to keep management convinced of the effectiveness of their bargaining weapon. Second, there appears to be the presumption in some of the literature that a breakdown of negotiations cannot occur if the two parties are "rational," so that one might argue that any breakdown of negotiations is due to the fact that the two parties are "irrational."[5]

It is not apparent how the propensity of either or both of the parties to (a) miscalculate the intentions of the other or (b) act irrationally would be systematically related to any of the conceptually observable variables in the system. Hence, the conventional bargaining theory approach is not very helpful in deriving implications about the frequency or duration of strikes.[6]

A. *An Alternative Bargaining Model*

A more fruitful—and more "realistic"—approach to the problem is to recognize at the outset that there are not two but three parties involved in labor-management negotiations: the management, the union leadership, and the union rank and file. This approach incorporates a set of institutional assumptions derived from the widely-accepted model of trade union behavior which is based on a separate analysis of the motivation of the union leadership and rank and file. By this view[7] the objectives of the leadership are: (1) the survival and growth of the union as an institution, and (2) the personal political survival of the leaders. These objectives are accomplished, in most part, by satisfying the expectations of the rank and file as well as possible. Even if the union is not democratic in a political sense, the leadership will in most cases respond to the desires of the membership for reasons of conviction. On the other hand, the leadership is aware of the possibilities of each bargaining situation, and it does more than merely represent the wishes of the rank and file. If the membership's expected wage increase is much greater than the management will agree to, the union leaders

[3] Explicit discussion of both of these issues is contained in Harsanyi [16] and Hicks [18].

[4] This is also the view expressed by Walton and McKersie [28, p. 56]. For a model which explicitly treats imperfect knowledge and a two-party learning process, see Cross [8].

[5] This position can be inferred from Harsanyi [17]. See also Bishop's comment on Harsanyi's paper in [5].

[6] Bishop states this explicitly: "It should be appreciated that neither Zeuthen's theory nor this one [Bishop's] involves a prediction of the frequency or duration of conflicts; each is really concerned only with the terms on which conflicts may be 'rationally' avoided" [4, p. 415].

[7] The following view of the nature of unionism is heavily dependent upon the position of Arthur M. Ross [24, esp. Chs. 1–3].

will attempt to convince the membership to be satisfied with a smaller increase. If they are unable to get the expected wage increase down to a sufficiently low level by the point of contract expiration, they face two alternatives: (1) signing an agreement which is less than the rank and file expects or (2) incurring a strike. If the leadership takes the first alternative, it faces the possibility that the contract will not be ratified by the membership and/or charges that they have "sold out" to management. The result will be internal union dissension and a decline in the political appeal and power of the leadership, both of which are antithetical to the basic objectives of the union leadership. The second alternative, although actually contrary to the membership's best interests, is preferred to the first by the leadership. Under strike conditions the leadership may at least appear as adversaries against management in a crusade which may even raise their political "stock" and will unify the workers. The outbreak of a strike, however, has the effect of lowering the rank and file's expectations due to the shock effect of the firm's resistance and the resultant loss of normal income. After some passage of time the leadership feels that the minimum acceptable wage increase has fallen to a level at which it can safely sign with management, and the strike ends.[8]

It is now possible to employ this essentially political model of the function of a strike to examine the firm's choice between giving in to the last union demand, which is the wage increase the rank and file finds acceptable as of the date of contract expiration, and "taking a strike" in order to obtain a lower settlement. The negotiated wage increase[9] which is acceptable to the union rank and file is

$$y_A \equiv \Delta W/\hat{W}, \tag{1}$$

where $\hat{W}$ is the previous contract wage rate and ΔW is the absolute wage increase. By the reasoning of the preceding discussion y_A depends on the length of the strike, S, say

$$y_A = v(S). \tag{2}$$

The precise shape of v is a matter of conjecture and surely differs between collective bargaining situations, but one would suppose that in the typical case it appears as in Figure 1.[10] Here $y_0 = v(0)$ is the acceptable wage increase at the point of contract expiration and $y_* = v(\infty)$ is the wage increase which the union would not accept with even an indefinitely long strike. This decay function may be represented as

$$y_A = y_* + (y_0 - y_*)e^{-\tau S}.\text{[11]} \tag{3}$$

For expository purposes let us suppose that

[8] Notable practitioners in the area of collective bargaining have long recognized this aspect of the function of a strike. William Simkin, Director of the Federal Mediation and Conciliation Service, has stated that: "If it is a fact, as it appears to be in many situations, that the union membership is unwilling to accept the reasonably attainable results of negotiations and is more militant than responsible leadership, a strike may be necessary to drive home the 'facts of life' " [26].

[9] Two comments are called for at this point. First, the analysis in the the text has been constructed on the assumption that bargaining takes place over the amount of a wage increase, not over the level of wages. This accords with the institutional literature with respect to collective bargaining, although this assumption could easily be changed. Second, nothing has been said in the text about bargaining over nonwage items, e.g., fringe benefits, and items relating to union functions such as company payment of shop stewards and dues check-off schemes. It is assumed implicitly throughout the text that these items have monetary equivalents and are imputed to the contracted wage.

[10] This is similar, but by no means identical, to Hicks's "union resistance curve" [28, p. 143].

[11] A more formal justification of the precise form of this equation may be obtained by assuming that under strike conditions the typical union member reduces his acceptable wage increase by some fraction of the difference between the currently acceptable wage increase and the lowest increase acceptable under any conditions, that is $\dot{y}_A = -\tau(y_A - y_*)$. This may be interpreted as a learning function, where the purpose of a strike is to set it in motion. Integration of this learning function gives equation (3) in the text. It is interesting to note that y_* may have any sign so that strikes to prevent wage decreases are by no means ruled out. The case $y_* \geq 0$ may be taken, however, as the more typical situation of downward wage rigidity.

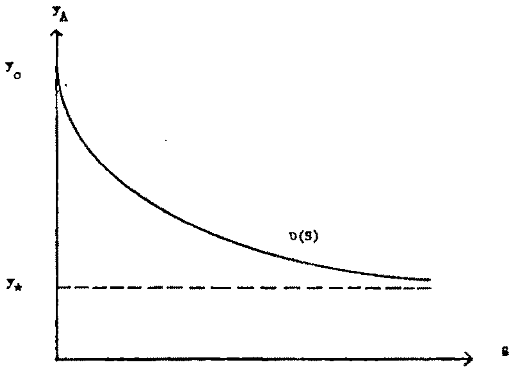


FIGURE 1

the typical firm is aware of the parameters of this relation and that it expects to produce a fixed output with the same technology to sell at the same price into the indefinite future. The profit level in each time period is

$$(4) \qquad \pi = \alpha P - \beta W - H,$$

where P is product price, H is the level of fixed production costs, and W is the negotiated wage rate. The latter may be rewritten from (1) as

$$(5) \qquad W = \hat{W}(1 + y_A).$$

The present value of the future profit stream is

$$(6) \qquad V = \int_0^\infty \pi e^{-rt}dt,$$

which may be written, after substitution of (3) into (5) and the result into (4), as

$$(7) \quad V = \int_S^\infty [\alpha P - \beta\hat{W}(1 + y_* + (y_0 - y_*)e^{-rS})]e^{-rt}dt - \int_0^\infty He^{-rt}dt.$$

Upon integration (7) becomes

$$(8) \quad V = [\alpha P - \beta\hat{W}(1 + y_* + (y_0 - y_*)e^{-rS})]\frac{e^{-rS}}{r} - \frac{H}{r},$$

which depends only on S, the length of the strike.[12] The firm that maximizes V has the choice of agreeing to y_0 and avoiding a strike or of rejecting y_0 and incurring a strike which will result in a lower wage increase. In effect, the firm must weigh the effect on profits of strike costs against the possibly lower wage costs which can be expected to accompany a strike. The firm maximizes V by not agreeing to y_0 and incurring a strike only if (i) $dV/dS=0$ and (ii) $d^2V/dS^2<0$ for some positive S; otherwise $S=0$ and $y_A=y_0$. Differentiating (8) and solving for S one obtains

$$(9) \quad S = -\frac{1}{\tau} \times \ln\left[\frac{\alpha P - \beta\hat{W}(1 + y_*)}{\beta\hat{W}\left(1 + \frac{\tau}{r}\right)(y_0 - y_*)}\right],$$

and the second order condition is satisfied when $y_0 > y_*$, which is true by assumption.[13] It follows from (9) that for a strike

[12] Thus far we have assumed that, even though the union-management relationship will continue indefinitely, contract negotiations take place only once. Without this assumption contract duration becomes a bargaining issue and expectations must be introduced explicitly into the analysis. Unfortunately, this involves complications all out of proportion to the authors' purpose in this paper. For an explicit justification of the assumption in the test, however, see Bishop [4, pp. 416–17].

[13] Needless to say, it is not necessary that firms actually make calculations such as those outlined in the text, only that they act as if such calculations had been made. On the other hand, there is some casual evidence to suggest that some firms explicitly engage in a maximization process similar to that noted above. The following quotations refer to the United Auto Workers-Ford Motor Company strike and may serve as an example of this casual evidence. They are taken from the *Ford Motor Company Report to Stockholders*, November 1967. "We are convinced that, in this situation, the UAW leadership concluded that no realistic settlement could be reached and ratified without a strike. . . . Given these difficult conditions, we believe the settlement we reached is a realistic one, even though it is higher than desirable. . . . A longer strike would have raised strike costs out of proportion to any resulting improvement in the outcome. In short, we believe the settlement represents the lowest possible combination of strike costs and settlement costs to the Company and the country."

to occur, $S>0$, it must be true that

$$y_0 > \frac{\alpha P - \beta \hat{W}\left(1 - \frac{\tau}{r} y_*\right)}{\beta \hat{W}\left(1 + \frac{\tau}{r}\right)}. \tag{10}$$

This inequality is more likely to be satisfied, *ceteris paribus*, the greater are y_0 and τ and less likely the greater are P, α/β (average product per worker, which is inversely related to the ratio of the wage bill to total cost), r, and y_*. By using (4) it is possible to rewrite (10) as

$$y_0 > \frac{\hat{\pi} + H + \frac{\tau}{r}\beta \hat{W} y_*}{\beta \hat{W}\left(1 + \frac{\tau}{r}\right)}, \tag{11}$$

which implies that a strike is less likely to occur over time in a given firm or industry the greater is the previous profit level relative to the previous wage bill.

Although the above model is a very oversimplified view of the collective bargaining relationship, it does provide predictions concerning the probability of a strike's occurrence and the expected duration of such a strike. Without going into great detail, the following remarks seem appropriate.

1. It is possible to increase the realism of this model substantially without drastically altering its implications. Introduction of the possibility of employment effects from wage increases, the fact that contracts are not negotiated once but at discrete intervals into the future, and the inventory position of the firm would all tend to increase the realism of the process. These extensions are unnecessary at this point, however, because the simple form of the present model is quite adequate for the purposes of estimation considered in the next section.

2. The present approach has the advantage of providing a determinate solution to the bargaining problem in the single (but important) case of union-management negotiations. Such a solution is possible only because widely held views about the institutional behavior of the parties involved is explicitly considered. In this case it is assumed that only one party, management, can realistically vary its wage offer. The union leadership, which maximizes its utility by acting in accord with the expectations of the rank and file, must act to represent the union membership's wage demands. On the assumption that firms maximize the appropriately discounted present value of the future profit stream, it is seen that the basic function of the strike is as an equilibrating mechanism to square up the union membership's wage expectations with what the firm may be prepared to pay. For completeness the analysis should probably include a fourth party, stockholders, in so far as management has a separate, self-serving motivation similar to that exposited by Williamson [30]. We feel, however, that for the present partial analysis we may safely assume that in the typical case management actions coincide with stockholder interests.

3. The above model has implications for variables other than the frequency and duration of strikes and the rate of change of money wages. Suppose, for example, that over a period of time wage changes remain below what the rank and file union membership desires, perhaps because of moral suasion via the Presidential wage guideposts. The analysis of this paper predicts that in such a case there will be an increase in the number of contracts which fail membership ratification and that there will be an increase in internal union dissension. George Perry has recently argued [22] that wage changes have been smaller since 1962 than would have been expected on the basis of the experience of the 1950s.

Interestingly enough, there has been a substantial increase in contract ratification defeats since the point at which Perry dates the initial overprediction.[14] Further, this period corresponds to a widely known rash of rank and file rebellion against union leadership.[15]

B. *An Operational Formulation of the Bargaining Model*

From the analysis of a typical firm's choice between incurring and not incurring a strike it was concluded that the parties were less likely to agree prior to conflict the greater the acceptable wage increase (y_0) and the speed at which the membership's expectations are reduced during a strike (τ); the parties are the more likely to agree the greater is the ratio of the pre-agreement profit level to the wage bill (π^*), the firm's discount rate (r), and the minimum acceptable wage increase (y_*). Only one of these variables, π^*, has an obvious empirical counterpart, so further hypotheses must be provided to relate the other variables to observable phenomena.

It seems plausible to argue that although τ, r, and y_* may vary between industries and regions because of different institutional arrangements, they will change only slowly through time. For example, τ should depend upon the size of benefits generally paid out of strike funds, how much unemployment compensation may be paid strikers, etc.; all of which are institutionally determined.[16]

Aggregating across firms and assuming a linear relationship in the relevant ranges of the variables, the preceding discussion suggests the following preliminary specification:

$$(12) \quad S_t' = \beta_0 + \beta_1 T + \beta_2 y_{0t} + \beta_3 \pi_{t-1}^*,$$

where S_t' is the probability of a strike in period t, and T indexes the passage of time. We expect $\beta_2 > 0$, $\beta_3 < 0$, $\beta_0 > 0$, since some strikes take place for institutional reasons, and $\beta_1 < 0$, since the number of institutional strikes has been steadily declining.[17]

Intuitively, one would expect y_{0t} to depend negatively on the unemployment rate, u_t. First, when unemployment is low the typical worker has the opportunity to move to a higher-paying job. Since the costs of movement may be substantial, however, he will first try to increase his wages in his present job and this will tend to increase y_0. Second, the leadership will be less likely to try to reduce y_0 when unemployment is low because the employment effects of a large wage increase will have little effect on their political stature, and sizeable strike funds may replace part of the worker's lost income. Finally, during periods of low unemployment there will be decreased opposition among the rank and file to a militant course of action since there will be part-time job opportunities for potential strikers.

[14] Perry uses the extent of overprediction of actual wage rate changes by a Phillips-type curve for the aggregate manufacturing sector to establish the dating of guidepost effectiveness. It is easy to see from his Table I [22, p. 899] that there is an increasing tendency for overprediction in the period from late 1962 through early 1966 (the last period presented). The following fiscal-year (July through June) data on the percentage of total joint-meeting Federal Mediation and Conciliation Service cases involving contract rejections match up very nicely: 1964–8.7 percent, 1965–10.0 percent, 1966–11.7 percent, 1967–14.2 percent. See Simkin [26].

[15] See, for example, Murray Gart [13]. Although most of this discontent has been manifest at the local level, it has also shown up in political turnover at the highest levels.

[16] A particular source of difficulty is encountered if y_* is a function of y_0, say $y_* = y_*(y_0)$. Even in this situation, however, the qualitative content of the model is retained so long as $dy_*/dy_0 < 1 + r/\tau$.

[17] The major reasons for this secular decline in strike activity have been enumerated succinctly by David Cole: "It has been possible over the years to all but eliminate two of the three major causes of strikes by resort to other means: disputes over recognition are now largely resolved by means of the election, and grievances by means of voluntary arbitration" [7, p. vii]. Data on strike activity by cause show that the number of strikes over union organization declined secularly from 839 in 1952 to 751 in 1957 to 582 in 1962. See [33, D-722].

A second determinant of y_0 should be a moving average of previous changes in real wages,

$$y_{0t} = \alpha_1 + \alpha_2 \sum_{i=0}^{M} \mu_i \Delta R_{t-i}. \tag{13}$$

Intuitively we would expect $\alpha_1 > 0$, $\alpha_2 < 0$, and the μ_i to have an inverted U-shape. That is, we would expect that when real wages have been increasing rapidly y_0 would be low. A more formal justification for these expectations is as follows: Suppose that y_0 depends positively on the difference between the expected long-run increase in real wages, ΔR^L, and the currently anticipated increase in real wages, ΔR^A, i. e.,

$$y_{0t} = \gamma_1[\Delta R_t^L - \Delta R_t^A]. \tag{14}$$

Suppose further that ΔR^L is composed of a very long-run constant increase ("workers always want more") and a moving average of previous real wage changes:

$$\Delta R_t^L = (1 - \sigma)V + \sigma \sum_{i=0}^{N} \delta_i \Delta R_{t-i}, \tag{15}$$

where $0 < \sigma < 1$, $\Sigma \delta_i = 1$, and V is the very long-run component. $\Delta R^A{}_t$ is presumably determined solely by a moving average of previous real wage changes:

$$\Delta R_t^A = \sum_{i=0}^{K} \lambda_i \Delta R_{t-i}, \tag{16}$$

where $\Sigma \lambda_i = 1$. Substitution of (15) and (16) into (14) gives:

$$\begin{aligned} y_{0t} &= \gamma_1(1 - \sigma)V \\ &\quad + \gamma_1 \sum_{i=0}^{M} [\sigma\delta_i - \lambda_i]\Delta R_{t-i}. \end{aligned} \tag{17}$$

Setting $\gamma_1(1-\sigma)V = \alpha_1$, $-\gamma_1 = \alpha_2$, and $(\lambda_i - \sigma\delta_i) = \mu_i$ gives precisely the form of (13). Note also that if the λ_i and δ_i have the usually assumed exponentially decaying form, then their difference will have an inverted U-shape.

A final determinant of y_{0t} should be profits. If the firm's profit level has been high in recent periods, the typical union member may feel that he deserves a larger wage increase. Also, the motivation of the leadership to attempt the task of persuading the membership to be content with a lower settlement will be diminished. Hence, high profit levels will have the effect of raising y_0 to some extent.

Combining the above hypotheses about the determinants of y_0, and assuming that the effect of profits on y_0 can be represented as $\alpha_4 \pi_{t-1}^*$, gives:

$$\begin{aligned} y_{0t} &= \alpha_1 + \alpha_2 \sum_{i=0}^{M} \mu_i \Delta R_{t-i} \\ &\quad + \alpha_3 u_t + \alpha_4 \pi_{t-1}^*. \end{aligned} \tag{18}$$

After substitution of (18) into (12) we have the following estimating equation:

$$\begin{aligned} \dot{S}_t &= A + B_1 \sum_{i=0}^{M} \mu_i \Delta R_{t-i} + B_2 u_t \\ &\quad + B_3 \pi_{t-1}^* + B_4 T + \epsilon_t, \end{aligned} \tag{19}$$

where A and the B_i are implicitly defined above and where ϵ_t is a disturbance term. On the basis of previous arguments we expect $A > 0$, $B_1\mu_i < 0$, $B_2 < 0$, and $B_4 < 0$. Since $B_3 = \beta_2\alpha_4 + \beta_3$, and $\beta_2\alpha_4 > 0$, $\beta_3 < 0$, the sign of the coefficient on profits in equation (19) is indeterminant. Although management is more likely to give in when previous profits are high, the union is also likely to increase its demands. Hence, it is not clear whether the net effect of an increase in profits will be to increase, decrease, or have no appreciable effect on the probability of occurrence of a strike. The specification of (19) concerning the effect of previous changes in real wages is based on the implicit assumption that workers view money wage changes and price changes as the reverse of each other. It is, of course, possibile that this is not so—complete money illusion being an extreme exception

—so this assumption is tested in the next section.

II. *Empirical Results*

A. *Specification and Estimation Problems*

If time-series observations were available on the number of strikes which begin in any quarter, S_t, and the number of contract expirations in any quarter, N_t, then S'_t could be set equal to S_t/N_t and (19) could be estimated directly. Although there are quarterly data for S_t, there are only limited surveys for N_t.[18] In essence, the difficulty faced is that even though equation (19) is derived from a model where the dependent variable is the probability of occurrence of a strike, it will be necessary to estimate it with data on only the frequency of occurrence of strikes.

Lacking the necessary observations on N_t, we are forced to make some plausible assumption about how it varies.[19] One possible hypothesis is that $N_t = n$, where n is a constant. Given the near-plateau in union membership reached in 1952 [27], and the increased tendency toward multiple-employer bargaining, this assumption does not seem implausible with respect to the *annual* number of contract expirations.[20] In order to deal with quarterly data, however, we must at least recognize that there is a strong seasonal influence in contract expirations. There are, of course, important economic reasons why this should be so. First, trade unions in the areas of the economy where inclement weather affects production and employment will always try to gear contract expirations to periods when the effects of a strike are least likely to be nullified by the fact that production would not have taken place anyway. Second, most trade unions will try to avoid contract expirations in periods when the demand for current income is high (the winter holidays, for example) and to obtain contract expirations in periods with the fewest paid holidays. Both of the above considerations suggest that the fall and winter quarters will contain fewer contract expirations and, *ceteris paribus*, fewer strikes. As a working assumption, therefore, we set

$$N_t = \sum_{j=1}^{4} \Phi_j N_{jt},$$

where Φ_j is the (constant) number of contract expirations in the jth quarter of all years and N_{jt} is a dummy variable set equal to one in the jth quarter of the year and zero otherwise. Substituting S_t/N_t for S'_t in (19) and multiplying both sides by

$$\sum_{j=1}^{4} \Phi_j N_{jt}$$

gives

$$(20)\quad S_t = A \sum \Phi_j N_{jt} + \sum \Phi_j N_{jt} X_t B + \sum \Phi_j N_{jt} \epsilon_t,$$

where for notational convenience the four independent variables in (19) have been replaced by the row vector X_t and B is the

[18] The data on S_t, for example, cover strikes involving more than six workers, while the limited data on N_t cover negotiations involving 1,000 or more workers. There is, of course, the further problem that some institutional strikes do not take place at a time of contract expiration.

[19] Needless to say, if we were willing to proceed by simply regressing the frequency of strike activity on any number of intuitively relevant independent variables, none of the above assumptions would be explicitly introduced. The use of such an *ad hoc* approach, however, would imperil any attempt to understand the mechanism which underlies the determination of strike activity.

[20] Some casual evidence for this assumption in the period 1963–67 is contained in Simkin [26]. The implication of the argument in the text is that the union always has its way on the seasonal aspect of disputes over contract expiration, which is not fully consistent with the caveat in footnote 16. In fact, the seasonal pattern implicitly suggested for strike activity has been observed over a long period of time. See, for example, Dale Yoder [31]. Our argument is simply that the precise period of the year in which a strike takes place is most generally under the control of the union, even if the point of contract expiration is not.

column vector of coefficients on these variables.

Since the Φ_j are unknown, the simplest procedure for obtaining an unbiased estimator for (20) would be to use a separate relationship for each quarter of the year. Unfortunately, the number of variables which will be needed on the right-hand side of (20) is too large to make this solution feasible with the length of the time series available. Therefore, as an admittedly rough approximation to (19) and (20), we have specified the following equation:

$$(21)\quad S_t = A\Phi_1 N_{1t} + A\Phi_2 N_{2t} + A\Phi_3 N_{3t} + A\Phi_4 N_{4t} + B_1' \sum_{i=0}^{M} \mu_i \Delta R_{t-i} + B_2' u_t + B_3' \overset{*}{\pi}_{t-1} + B_4' T + \epsilon_t$$

where the N_{jt} are a set of seasonal dummies and ϵ_t' is an error term.[21] It is not difficult to show, under the usual assumptions about the way in which the independent variables in (21) are generated [14, pp. 268–69], that the least squares estimators of $B_1'\mu_i$, B_2', B_3' and B_4' are consistent and asymptotically unbiased estimators of $\frac{1}{4}\Sigma\Phi_j(B_1\mu_i)$, $\frac{1}{4}\Sigma\Phi_j B_2$, $\frac{1}{4}\Sigma\Phi_j B_3$, and $\frac{1}{4}\Sigma\Phi_j B_4$. That is, our estimates of the coefficients in equation (21) may be interpreted as the *mean response*, averaged across quarters, of a stimulus from the independent variable. Since each of the Φ_j must be positive, all of the previous predictions about the signs of the coefficients in equation (19) hold for equation (21).

Before turning to estimation and testing of equation (21), one final difficulty must be resolved. The lag coefficients on ΔR_{t-i} pose two estimation problems. First, we have a priori reason to suppose that these coefficients will not have the familiar exponential-decay form. Second, we expect a lag distribution on ΔR_{t-i}, but not on the other variables in the equation.[22] As an alternative to the standard technique, therefore, we have used the straightforward procedure suggested by Shirley Almon [1] for estimating equation (21). On the assumption that the lag distribution can be approximated by a polynomial, Lagrangian interpolation polynomials are used to create moving averages of the independent variable. These moving averages are then introduced into an ordinary regression equation and their coefficients (and hence the implied coefficients of the lag distribution) estimated. In the results that follow we have used a third degree polynomial approximation and constrained the lag distribution to assume zero values at the beginning and at a finite lag.[23]

B. *The Results*

Our discussion of empirical results is divided into three sections. First, we discuss the general implications of the estimated version of equation (21). Second, we modify the estimating equation to allow for the possibility of money illusion on the part of workers. Finally, we test the stability of the preferred equation and consider the effect of some important institutional changes on the volume of aggregate strike activity.

1. The initial results of fitting equation (21) are reported in Tables 1 and 2 under the rubric of equations (21a) and (21b),

[21] If the disturbance term in (19) is homoscedastic, then the disturbance term in (21) is heteroscedastic because its variance moves systematically with the quarter of the year being considered. Namely, if var $(\epsilon_t) = C$, a constant, then var $(\epsilon_t') =$ var $(\Phi_j\epsilon_t) = \Phi_j^2 C$. If one views S_t' in (19) as a sample proportion, however, the procedure used to obtain (21) suggests that ϵ_t' will be (nearly) homoscedastic. Since this becomes an empirical issue, we have tested the estimated version of (21) for unequal residual variances to see if this difficulty is damaging to our results. See footnote 24 below.

[22] For a discussion of the estimation problems raised by these difficulties see Griliches [15].

[23] This means that two "Almon variables" were entered into the regression to estimate the lag distribution on real wage changes. It should be noted, however, that the results were generally insensitive to either the degree of the polynomial or the constraints placed on the lag coefficients.

TABLE 1—ESTIMATED REGRESSION COEFFICIENTS AND RELATED STATISTICS[a,b] FOR THE VARIABLES IN EQUATIONS 21a, 21b, 21c, AND 21d

Equation	U_t	$\Sigma\Delta R_{t-i}$	$\Sigma\Delta W_{t-i}$	$\Sigma\Delta P_{t-i}$	π^*_{t-i}	N_1	N_2	N_3	T	C	LG	R^2	$\hat{R}^2$	DW	SEE
(21a)	−123.0	−62.2	—	—	1.6	213.6	594.8	457.9	−2.2	1519.8	—	.938	.820	1.44	75.9
	(13.1)	(12.9)			(136.7)	(30.8)	(28.4)	(27.9)	(0.7)	(170.0)					
(21b)	−123.2	−62.2	—	—	—	213.7	594.8	457.9	−2.2	1521.7	—	.938	.820	1.44	75.2
	(9.4)	(12.3)				(28.7)	(27.4)	(27.5)	(0.7)	(70.4)					
(21c)	−132.6	—	−80.6	64.4	—	227.3	602.4	459.4	−2.8	1663.8	—	.941	.828	1.52	75.0
	(11.8)		(24.7)	(14.2)		(30.2)	(27.8)	(27.4)	(1.1)	(168.4)					
(21d)	−135.3	−62.9	—	—	—	225.7	598.7	460.5	−2.3	1570.4	87.8	.946	.843	1.61	70.7
	(9.8)	(11.5)				(27.3)	(25.8)	(25.8)	(0.6)	(68.4)	(30.9)				

[a] R^2 is the coefficient of determination about the overall mean, $\hat{R}^2$ is the coefficient of determination about the quarterly means, *DW* is the Durbin-Watson statistic, and *SEE* is the Standard Error of Estimate for the regression equation.

[b] Estimated standard errors of the estimated regression coefficients are in parentheses under the relevant coefficients.

Sources: The civilian unemployment rate, u_t, is a quarterly average of the monthly rates published in Table A-6 of the *Monthly Labor Review*, U. S. Bureau of Labor Statistics. The Consumer Price Index is obtained from D-1 of the *MLR.* Our wage rate is an average, weighted by relative 1957 production worker employment, of average hourly earnings in mining, construction, and manufacturing, and the data are obtained from *Employment and Earnings* volumes, U. S. Bureau of Labor Statistics. None of these variables is seasonally adjusted. Our profits variable is the ratio of Corporate Profits after tax, excluding Inventory Valuation Adjustment, to Total Compensation. The source of these data is various issues of the *Survey of Current Business*, Table 3, U. S. Dept. of Commerce. Finally, the number of strikes beginning in each quarter are obtained in Table E-1 of various issues of the *MLR.*

TABLE 2—ESTIMATED LAG COEFFICIENTS AND THEIR STANDARD ERRORS FOR EQUATIONS 21b AND 21c

Lag	(21b) $\Sigma\Delta R_{t-i}$	(21c) $\Sigma\Delta W_{t-i}$	$\Sigma\Delta P_{t-i}$
0	−4.0 (1.6)	− 7.0 (2.6)	4.5 (2.0)
1	−6.9 (2.3)	−11.4 (4.0)	7.5 (2.9)
2	−8.6 (2.4)	−13.4 (4.3)	9.3 (3.0)
3	−9.5 (2.1)	−13.5 (4.1)	10.0 (2.6)
4	−9.4 (1.9)	−12.2 (3.7)	9.8 (2.2)
5	−8.6 (1.9)	−10.0 (3.6)	8.7 (2.0)
6	−7.1 (2.1)	− 7.2 (3.5)	7.1 (2.1)
7	−5.2 (2.1)	− 4.3 (3.2)	4.9 (2.1)
8	−2.7 (1.5)	− 1.8 (2.2)	2.5 (1.5)

and the lag distribution coefficients for equation (21b) are charted in Figure 2. The following symbols are used in these tables:

u_t = the civilian unemployment rate,
$\Delta R_t = \Delta W_t - \Delta P_t$,
ΔW_t = the annual percentage rate of change of money wages,
ΔP_t = the annual percentage rate of change of consumer prices,
π_t^* = the ratio of corporate profits after taxes to total compensation,
N_{jt} = seasonal dummies for first, second, and third quarters,
T = time in quarters,
C = constant term.

The period of fit is 1952I-1967II, which is consistent with our assumption that the annual number of contract expirations has been relatively constant since 1952.

As can be seen from Tables 1 and 2, the results provide strong support for the hypotheses advanced in Section I. In equation (21a) the coefficients of each of the independent variables, except that for profits, are highly significant.[24] Although several different measures of profits and moving averages of profits were tried, none produced results substantially different from those reported in Table 1. Our tentative conclusion is that the *net* effect of profits *on strike activity* is small. In equation (21b) π^*_{t-1} has been deleted. The independent variables in this preferred equation explain about 94 per cent of the variance of the dependent variable about its overall mean, and about 82 per cent of the variance of the dependent variable about its quarterly means. The standard error for

[24] Since there is some question about the appropriateness of the usual statistical tests in the case where the disturbance variances are unequal (see footnote 21), we have applied Bartlett's well-known test for unequal variances to these data. See Bennett and Franklin [3] concerning the computational procedures of this test. For the number of degrees of freedom in each of the quarterly groupings, B, the test statistic, is satisfactorily approximated by the χ^2 distribution with 3 degrees of freedom. In this case, $B = 2.01$, and we cannot reject the hypothesis of equal quarterly residual variances at even the .25 level.

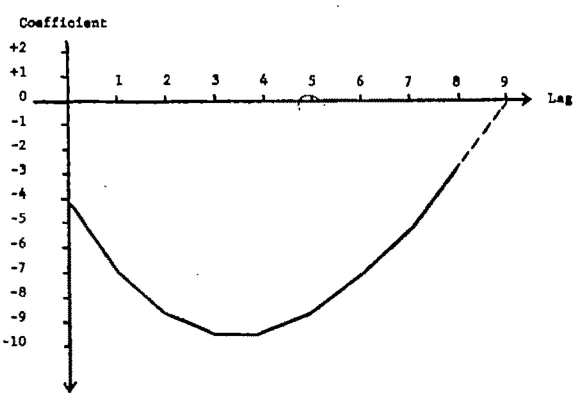


FIGURE 2—LAG DISTRIBUTION COEFFICIENTS ON REAL WAGE CHANGES

this equation is a remarkably low 75 strikes, as against a mean of about 980 strikes per quarter. A glance at Figure 2 suggests that the lag distribution on real wages is of the shape predicted. A steady-state decline of one percentage point in the rate of change of real wages is associated with an increase of about 62 strikes per quarter. The seasonal dummies confirm the hypothesis that strike activity is much heavier in the spring and summer quarters than in the fall and winter quarters.[25] Finally, a decline of one percentage point in the civilian unemployment rate is associated with an increase of about 123 strikes per quarter.

2. The results for equation (21c) reported in Tables 1 and 2 incorporate the hypothesis that the rates of change of money wages and prices are not mirror images in their effects on aggregate strike activity, i.e., we allow for the possibility that the rate of change of money wages has a more (or less) important effect on strike activity than the rate of change of prices. It is clear from these results that the effect of money wage changes on strike activity is somewhat greater than the effect of price changes. The differences in the effects of these variables, however, is not very substantial. More formally, we may test the null hypothesis that

$$\sum_{i=0}^{9} \mu_i^w + \sum_{i=0}^{9} \mu_i^p = 0$$

by forming the ratio

$$t = \frac{\sum \hat{\mu}_i^w + \sum \hat{\mu}_i^p}{\text{var} \quad (\sum \hat{\mu}_i^w + \sum \hat{\mu}_i^p)^{\frac{1}{2}}},$$

where the $\hat{\mu}_i^w$ are the estimated lag distribution coefficients on money wage changes and the $\hat{\mu}_i^p$ are the estimated lag distribution coefficients on price changes. In this case $t = -1.07$, which clearly is not significant at conventional test levels. We may also test the null hypothesis $\mu_i^w = -\mu_i^p$ ($i = 0, \cdots, 9$) with an F-ratio. In this case $F(2, 49) = 1.09$, which also is not significant at conventional test levels.[26] We tentatively conclude that wage and price changes act essentially as mirror-images with respect to aggregate strike activity.

3. Since the volume of strike activity has heretofore been associated with any number of unstable causal factors, there may be some question about the general stability of an equation like (21b) over time. In order to test for the possible instability of this equation we have arbitrarily divided the sample period in half and performed the standard test of the null hypothesis that the parameters of equation (21b) are identical for the two time periods. In this case $F(8, 46) = 1.40$, which clearly is not significant at conventional test levels. We conclude that there is little evidence to suggest that this relationship is unstable.

[25] The coefficients on the seasonal dummies and the constant term in equation (21b) allow us to estimate $A\Sigma\Phi_j$ and $A\Phi_j (i = 1, \cdots, 4)$. Hence, they provide estimates of

$$\frac{A\Phi_j}{A\Sigma\Phi_j} = \frac{\Phi_j}{\Sigma\Phi_j} (j = 1, \cdots, 4),$$

i.e., the percentage of total annual contract expirations which take place in each quarter. These estimates are, from the first to fourth quarters: 23.6 per cent, 28.9 per cent, 26.9 per cent, 20.4 per cent. Simkin [26] provides the number of "active" Federal Mediation and Conciliation Service cases closed each month in 1966, which, assuming a short lag from beginning to closure of a case, should be a good proxy for the "typical" number of contract expirations per month. Assuming a mean lag of one month for closures, these independent data provide the following estimates of the percentage of total contract expirations which take place in each quarter: 22.8 per cent, 32.3 per cent, 25.6 per cent, 19.2 per cent. These latter estimates are strikingly close to those obtained from equation (21b), and lend additional credence to our argument that the seasonality in strike activity is due to a seasonality in contract expirations.

[26] The first test described in the above paragraph requires computation of the estimated variance of $\Sigma\hat{\mu}_i^w + \Sigma\hat{\mu}_i^p$, which can easily be worked out as a linear combination of the variances and covariances of the "Almon variables" in the regression equation. The second test is a straightforward application of some of the results in Chow [6].

Finally, we have estimated equation (21d) to allow for the possibility that passage of the Landrum-Griffin Act in 1959 has had a positive impact on the amount of aggregate strike activity. It has been argued that this law, which is designed to regulate the internal affairs of trade unions in order to ensure "union democracy,"[27] has had the effect of: (a) increasing the militancy of union leaders as a response to the implicit encouragement the law gives to the growth of dissident groups within the union, and (b) making the leadership more sensitive to the "less responsible" wage demands of the union members.[28] Adding a dummy variable, LG, to the estimating equation to test for the effect of the Landrum-Griffin Act gives the results reported as equation (21d) in Table 1. The standard error, Durbin-Watson Statistic, and R^2 are all improved by this modification. The coefficient of LG suggests a modest, but significant increase of about .88 strikes per quarter over the pre-Landrum-Griffin period.[29]

III. *Conclusions and Implications*

Although we are not firmly wedded to the precise estimates presented in this paper, it seems that the aggregate level of strike activity is behaviorally related to the degree of tightness of the labor market and previous rates of change of real wages. Among the implications of our analysis are the following:

1. The incorporation of a widely accepted set of assumptions about the behavior of trade unions into the traditional theory of the firm produces a straightforward solution to the outcome of union-management bargaining which lies within the corpus of conventional economic reasoning. Although conventional bargaining models are based on assumptions which do not seem to represent the institutional framework of union-management negotiations, this could be excused if they provided refutable predictions about observable behavior, but they do not. The simple formulation of this paper, which specifically acknowledges the ***three-party nature*** of collective bargaining, does yield refutable predictions, and they are found to be consistent with the data.

2. As was shown with the case of the Landrum-Griffin Act, even a simple version of the model described in this paper can be helpful in evaluation of the effects of changes in the institutional framework within which collective bargaining must function. Other institutional changes, e.g., the payment of unemployment compensation to strikers, could be investigated. Further, the model described in this paper provides a more explicit rationale for some of the work which has been done on explanations of the rate of change of aggregate money wages.[30]

3. Finally, the results have a number of implications for public policy with respect to wage determination in the unionized sector of the economy. First, policies which

[27] See [12] for details on the provisions of this law. It was signed in September of 1959 and most of its provisions were in effect within ninety days. Hence, in the following results we have assumed that the law began having an effect in 1959 IV and reached its full effectiveness linearly by 1960 IV.

[28] For arguments similar to these see Estey [11, pp. 57–59). In terms of the formulation in Section I, these arguments suggest that y_0 would generally be higher after passage of the Landrum-Griffin Act than before.

[29] To test for the possibility that we had confused the effect of the Landrum-Griffin Act with the effect of the Guideposts, we added another dummy variable to the equation. The coefficient of LG remained similar with respect to its standard error while the Guideposts coefficient was negative and less than half its standard error. In terms of the formulation in Section I, this negative effect is what would be predicted since the Guideposts supposedly lower y_0.

[30] For example, the studies by O. Eckstein and T. A. Wilson [10] and G. L. Perry [21] both include profits as determinants of money wage changes on the basis of casual bargaining model considerations, and J. D. Sargan [25] estimates a model which includes the level of real wages. In [2] a variable specifically measuring the volume of strike activity is found to have an effect on aggregate money wage changes.

are geared to induce labor leaders to convince their constituencies to be satisfied with "more reasonable" wage settlements are likely to result eventually in political turmoil within trade unions. It does not, therefore, seem likely that such a policy can continue for long without encouraging the growth and power of more militant leadership and a subsequent decline in the effectiveness of the original policy. Second, in addition to the well-known tradeoff between wage changes and unemployment there seem also to be tradeoffs between unemployment, wage changes, and industrial strike activity. To the extent that the maintenance of industrial peace is a goal of public policy, this adds an additional dimension to the problem of maintaining full employment and stable prices.

References

1. S. Almon, "The Distributed Lag Between Capital Appropriations and Expenditures," *Econometrica*, Jan. 1965, *33*, 178–96.
2. O. Ashenfelter, G. E. Johnson, and J. H. Pencavel, "Trade Unions and the Rate of Change of Money Wages in the United States: 1914–1960," Industrial Relations Section, Princeton University, (mimeo), 1968.
3. C. A. Bennett and N. L. Franklin, *Statistical Analysis in Chemistry and the Chemical Industry*. New York 1954.
4. R. L. Bishop, "A Zeuthen-Hicks Theory of Bargaining," *Econometrica*, July 1964, *32*, 410–17.
5. ———, "Game-Theoretic Analyses of Bargaining," *Quart. Jour. Econ.*, Nov. 1963, *77*, 559–602.
6. G. C. Chow, "Tests of Equality Between Sets of Coefficients in Two Linear Regressions," *Econometrica*, July 1960, *28*, 591–605.
7. D. Cole, "Preface," in J. Dunlop and N. Chamberlain, eds., *Frontiers of Collective Bargaining*, New York 1967, pp. vii-ix.
8. J. G. Cross, "A Theory of the Bargaining Process," *Am. Econ. Rev.*, March 1965, *55*, 67–94.
9. J. Dunlop, *Wage Determination Under Trade Unions*. New York 1944.
10. O. Eckstein and T. A. Wilson, "The Determinants of Money Wage Changes in American Industry," *Quart. Jour. Econ.*, Aug. 1962, *76*, 379–414.
11. M. Estey, *The Unions*. New York 1967.
12. M. S. Estey, P. Taft, and M. Wagner, eds., *Regulating Union Government*. New York 1954.
13. M. Gart, "Labor's Rebellious Rank and File," *Fortune*, Nov. 1966, *74*, 150–53.
14. A. S. Goldberger, *Econometric Theory*. New York 1964.
15. Z. Griliches, "Distributed Lags: A Survey," *Econometrica*, Jan. 1967, *35*, 16–49.
16. J. C. Harsanyi, "Approaches to the Bargaining Problem Before and After the Theory of Games," *Econometrica*, April 1956, *24*, 144–57.
17. ———, "Notes on the Bargaining Problem," *So. Econ. Jour.*, April 1958, *24*, 471–76.
18. J. R. Hicks, *The Theory of Wages*. New York 1963.
19. C. C. Holt, "Job Search, Phillips' Wage Relation and Union Influence, Theory and Evidence," Univ. of Wisconsin Firm and Workshop Paper 6705, Dec. 1967.
20. F. S. O'Brien, "Industrial Conflict and Business Fluctuations: A Comment," *Jour. Pol. Econ.*, Dec. 1965, *73*, 650–54.
21. G. L. Perry, "The Determinants of Wage Rate Changes and the Inflation—Unemployment Tradeoff for the United States," *Rev. Econ. Stud.*, Oct. 1964, *31*, 287–308.
22. ———, "Wages and the Guideposts," *Am. Econ. Rev.*, Sept. 1967, *57*, 897–904.
23. A. Rees, "Industrial Conflict and Business Fluctuations," *Jour. Pol. Econ.*, Oct. 1952, *60*, 371–82.
24. A. M. Ross, *Trade Union Wage Policy*. Berkeley and Los Angeles 1948.
25. J. D. Sargan, "Wages and Prices in the United Kingdom," in P. E. Hart, G. Mills, and J. K. Whitaker, eds., *Econo-*

metric Analysis for National Economic Planning, London 1964, pp. 25–54.

26. W. Simkin, "Refusals to Ratify Contracts," *Lab. Rel. Reporter*, Nov. 16, 1967, 66LRR 234.
27. L. Troy, "Trade Union Membership, 1897–1962," *Rev. Econ. Stat.*, Feb. 1965, *47*, 93–113.
28. R. E. Walton and R. B. McKersie, *A Behavioral Theory of Labor Negotiations*. New York 1965.
29. A. R. Weintraub, "Prosperity vs. Strikes: An Empirical Approach," *Indus. Lab. Rel. Rev.*, Jan. 1966, *19*, 231–38.
30. O. E. Williamson, *Economics of Discretionary Behavior: Management Objectives in a Theory of the Firm*, Englewood Cliffs, N. J. 1964.
31. D. Yoder, "Seasonality in Strikes," *Jour. Am. Stat. Assoc.*, Dec. 1938, *33*, 687–93.
32. F. Zeuthen, *Problems of Monopoly and Economic Warfare*. London 1930.
33. U. S. Bureau of the Census, *Historical Statistics of the United States*. Washington, D.C. 1960.

A Spectral Analysis of Post-Accord Federal Open Market Operations

By VITTORIO BONOMO AND CHARLES SCHOTTA*

This paper applies spectral analysis to time series data on post-Accord Federal Reserve Open Market Operations (*OMO*), member bank reserves, and member bank free reserves to ascertain the effect of *OMO* on variations in reserve levels.[1]

The first section of our paper describes the data we used. The next section briefly outlines spectral analysis; no formal mathematical presentation of time series analysis is attempted since rather good compact treatments are readily available. However, since few economists are familiar with the details of this approach to time series analysis, we present the basics which are necessary for an understanding of our findings. In addition, some of the problems encountered in applying time series analysis to economic data are mentioned. The third section presents the statistical results of our analysis while the final section makes some interpretation of these results.

If one interprets the role of defensive open market operations as that of reducing short-term fluctuations in bank reserves which arise from sources such as changes in the public's demand for currency, changes in float, changes in foreign clearings, etc., then our analysis indicates that the Fed has had considerable success in achieving this goal during the post-Accord period. Our conclusions are much less secure if one questions whether this has been the goal of *OMO*. This, however, is not a difficulty peculiar to our study but is at least implicit in any evaluation of an entity where some objective function must be assumed.

I. *The Data*

The data we used in our analysis consist of week-to-week changes in the Federal Reserve System's holdings of U.S. government debt of all kinds beginning, as do the other series we use, with the week of April 4, 1951 and ending with the week of May 31, 1967. These and the other data, from various issues of the *Federal Reserve Bulletin*, are weekly averages of daily data and are not seasonally adjusted. This first series, week-to-week changes in the Fed's holdings of U.S. government securities, we call *G*. These

* The authors are at Virginia Polytechnic Institute. An earlier version of this paper was presented at the Econometric Society Summer meetings, Toronto, August 29, 1967. We are grateful to Philip Cagan, Thomas Cargill, C. W. J. Granger, A. O. Hughes, Thomas Mayer, and Clark Warburton for their comments on earlier drafts of this paper. J. G. Gurley and the referees of this journal made valuable suggestions. Dana Burtness provided computational and typing assistance. The Computer Centers of the University of California, Santa Barbara and Davis Campuses provided funds for computation. Most of the research was accomplished while Bonomo was on the staff of the University of California, Santa Barbara and Schotta was on the staff of the University of California, Davis, with financial assistance from its Institute of Governmental Affairs. We remain responsible for any errors.

[1] There are now several excellent reference monographs on the statistical theory underlying spectral analysis. For the general economist perhaps the best introduction to the use of spectral methods in economics is C. W. J. Granger and M. Hatanaka [10]. A very recent reference at a slightly more advanced level, although not so comprehensive and applications-oriented as [10] is G. S. Fishman [9]. For a thorough treatment of spectral theory, including applications drawn primarily from engineering and the physical sciences see N. S. Bendat and A. G. Piersol [1]. The most comprehensive treatment of spectral analysis yet in print is the excellent and very recent G. M. Jenkins and D. G. Watts [12].

changes are assumed to occur only as a result of open market operations.[2]

The data for reserves are week-to-week changes in member bank reserve levels and week-to-week changes in member bank free reserve levels. These are called, respectively, R and FR. These changes are assumed to be the result of all factors, including *OMO*, which can change reserve levels and free reserve levels.[3]

In our study we are interested in separating the changes in R and FR which are produced by G from those changes produced by the other factors influencing reserve levels. For this purpose, we construct two series of the same length as G, R, and FR, namely

$$R_t^* = R_t - G_t \tag{1}$$

which we call week-to-week changes in member bank reserve levels ***net*** of Federal Reserve open market operations and

$$FR_t^* = FR_t - G_t \tag{2}$$

which we call week-to-week changes in FR ***net*** of *OMO*. Thus we have five time series, each of 843 observations, as our basic data.

We assume in our analysis that the factors which influence R^* and FR^* are not functionally dependent upon G.[4] However, to the extent that open market operations are undertaken for defensive purposes, G will be functionally dependent upon factors affecting R^* and FR^*.

II. *A Sketch of Spectral Analysis*

Before describing the results of our analysis of these time series, it may be useful to present a brief survey of spectral analysis and to note some of the problems in the application of this statistical method—one which has proven so fruitful in certain physical sciences, such as electrical engineering, engineering geoscience, oceanography, and communications engineering—to economic time series.[5]

The essence of the spectral approach is to treat a time series as a signal produced by a physical stochastic process. This signal may contain many frequencies, not all of which necessarily contain information. The aim of the analysis is to examine the signal, classify its characteristics, and make inferences about the process which generated it (which presumably we cannot observe directly). This approach separates the signal (time series) into many different "frequencies." Then the relative amount of the total "energy" of the signal which is contained in each of the frequencies is cal-

[2] It is possible for the Fed's U.S. Government portfolio to change as a result of events other than an open market sale or purchase. For example, if the Fed holds some short maturity bills and allows them to run to cash, then the portfolio falls and the Treasury accounts fall by the same amount. Such events are quite rare, however, and an examination of changes in the System Open Market Account has convinced us that *OMO* represent by far the largest source of change in this portfolio.

[3] One source of change in R and FR which was of considerable interest to us is an autonomous change in the public's desire for currency and coin. An increase in currency outside all banks reduces member bank reserves dollar for dollar at the present time. Prior to the incorporation of vault cash in the reserve base, it was necessary to hypothesize a vault cash term in the banks' portfolio demand function such that there was a dollar-for-dollar replacement of cash which was withdrawn by the public. One other important source of change in reserves is the volume of float. We are presently attempting to expand our study to include an analysis of this very interesting phenomenon.

[4] We could see no a priori dependent connection between any of the other factors changing the level of reserves and open market operations, except possibly currency (C). In an earlier paper, V. Bonomo and C. Schotta [4], we reported on our investigation of the role of C as a dependent variable of current G as well as past values of G. Although we have found that we cannot reject the hypothesis that there is a dependent relationship, it is extremely small. Hence we could not see any necessity for a further adjustment of R^* and FR^* to reflect a further influence of G through C.

[5] The contents of this section and the discussion of cross-spectral methods below are intended only to present the computing formulae used so that the reader unfamiliar with spectral analysis can at least see what has been done. No attempt is made to present the properties of the estimators or to discuss the statistical theory underlying the method.

culated. If we have a time series sample $\{x_t, t = 1, \ldots, n\}$ which meets the requirement that

$$\text{(3)} \qquad \mathrm{Cov}\,(x_1, x_2) = \mathrm{Cov}\,(x_{1+h}, x_{2+h}),$$

then we decide the number of frequency components, m, into which we wish to separate the series.[6] The first step in securing the power spectrum estimate is to secure the auto-covariance of the series for $m+1$ lags

$$\text{(4)} \qquad C(i) = \frac{1}{n}\left[\sum_{t=1}^{n-(i-1)} x_t x_{t-(i-1)} - \frac{1}{n}\left(\sum_{t=1}^{n} x_t\right)\left(\sum_{t=1}^{n} x_t\right)\right].$$

where $i = 1, 2, \ldots, (m + 1)$.

This auto-covariance function then is converted to the estimates of the power spectrum by calculating a weighted auto-covariance

$$\text{(5)} \qquad C'(i) = [w(i)][C(i)]$$

for

$$i = 1, 2, \cdots, m + 1$$

where

$$\text{(6a)} \quad w(i) = 1 - 6\left(\frac{i}{m}\right)^2\left(1 - \frac{1}{m}\right)$$

for

$$1 \leq i \leq \frac{m}{2}$$

$$\text{(6b)} \qquad w(i) = 2\left(1 - \frac{1}{m}\right)^3$$

for

$$\frac{m}{2} + 1 \leq i \leq m + 1$$

[6] Selection of the number of frequency components, in addition to determining the "resolution," also determines the degrees of freedom for tests of hypotheses. See Jenkins and Watts [12] for a discussion of hypothesis testing. Bonomo and Schotta [3] contains tabulated confidence intervals for the most commonly used spectral windows.

The weighted auto-covariance is then subjected to a Fourier cosine transformation to form estimates of the power spectrum

$$\text{(7)} \qquad \hat{f}(i) = \sum_{j=1}^{m+1} \lambda C'(j) \cos \frac{(j-1)(i-1)\Pi}{m}$$

for

$$i = 1, 2, \cdots, m + 1$$

where

$$\lambda = 1 \qquad \text{for } j = 1; j = m + 1$$
$$\lambda = 2 \qquad \text{for } j = 2, 3, \cdots, m.$$

The auto-covariance can be recovered by means of an inverse transformation. The power spectrum provides a decomposition of the variance of a time series in the frequency domain since

$$\text{(8)} \qquad \sigma_x^2 = C(0) = \frac{1}{m}\int_{\omega=0}^{m} \hat{f}(\omega)d\omega$$

Economists have been practicing a crude form of spectral analysis (or variance decomposition in the frequency domain) for some time when they separate a time series, say pig iron production, into a "trend" component, a "cycle" component, a "seasonal variation" component, and, perhaps, an "irregular variation" component. These components, however, are themselves relatively broadband signals and can be used only to infer the grossest kinds of things about the physical process itself. For example, suppose we find a strong "cycle" component. This may be a very wide band and the "information" in this band may lie only in a narrow region with all other frequencies within the band containing only "noise."

In general, it is advisable to use as many lags (hence, as many frequency elements) as possible. If any useful confidence bands for the spectral density function are to be retained, the limitation on the number of lags is given by the ratio of

the number of data points to the number of lags which should not be less than 4. Writers on spectral analysis such as Jenkins and Watts [12, pp. 277–79] point out that the choice here is between "resolution," i.e., the detail involved in the estimate of the spectral density function (which is an increasing function of the number of lags), and the width of the confidence interval at the selected probability level. In practice, the only useful solution to the problem is to expand the number of data points. Occasionally, spectra have been crudely estimated with as few as 100 data points and 25–30 lags; however, the resulting spectrum is not only "coarse" but the confidence interval at, for example, the .95 probability level is likely to be wide.

This limitation on the number of frequency components (and the accompanying requirement for large amounts of data) is the most serious barrier to widespread use for analysis of economic time series. There are, after all, only some 80 or so postwar quarters which, even at the outside, allow only 25 or so frequency components.

Another serious limitation on the use of spectral analysis for economic time series lies in the necessity for stationary time series for the statistical theory results to be applicable. Most economic time series are subject to quite considerable time trends which appear as very low frequency cycles of dominant strength when estimated.[7]

There is a limitation on the use of spectral analysis for economic time series, not usually mentioned in treatments of the subject, which arises from the method of securing a digital data set. In the typical application of time series analysis in, say, engineering geoscience, the data consist of a continuous time series derived from, say, an instrument recording seismic noise in a particular location. Once the geophysicist had collected a record of sufficient length he would then be able to choose the interval for use for digitizing the data. In order to ascertain whether these data were sensitive to the interval he could choose one or more alternates to see if the characteristics of the time series differed between intervals. A simple example of the type of distortion which could arise because of an improperly chosen digitizing interval is provided when one samples a series with a fast cosine wave. The result may be a series exhibiting a slow cosine wave.

Most economic time series are reported with the same interval between data points (although in the case of monthly data even this is not true). However, the digitizing interval which is built into the data collection effort is usually fixed and constitutes a sample of points from the realization of a physical process, the characteristics of which are unknown. It is as if we decided to take the temperature at a certain location once a day at, say, noon. This interval would cause us to make inferences about the physical process, either in a descriptive sense or in a causative sense, which would probably be erroneous.

With respect to the time series with which we are working, weekly averages of daily data, we are attempting to obtain the daily data themselves to examine the resulting frequencies which are now above the .5 cycle per week frequency, the highest frequency we can presently estimate.[8]

[7] For discussions of methods of attacking this problem without transformations which in effect remove a portion of the information in the time series see M. B. Priestly [16] and the very interesting contribution by Lloyd Brown [8].

[8] This frequency, the "Nyquist," depends on the digital interval selected in the data collection (or, in the engineering data case, the process of digitizing a continuous time series sample). Bendat and Piersol [1, p. 291] recommended selecting a digital interval such that the Nyquist is twice the highest frequency of any analytical interest. Since our analysis concerns no fre-

III. *Our Results*

When we estimate the power spectrum for each series, we obtain the "energy" associated with the particular frequency for the series. We have estimated 157 frequencies for each series. Because of the variance decomposition in the frequency domain we have the record of the contribution of each of these frequency components to the total auto-covariance of the series. For convenience in description we have converted the spectral power for each frequency into the percentage of the summation of the spectral power for all frequencies.

$$Z(i) = \hat{f}(i) / \sum_{j=1}^{m+1} f(j) \tag{9}$$

where

$$i = 1, 2, \cdots, m + 1.$$

This measure, Z, then represents the percentage of the total variance of the series which is associated with each frequency.

In general (as with any signal) one is interested primarily in those frequency bands that contain the most "information" since the "message," as we usually think of it, is contained there. If we have a broadband signal with all the energy in only one narrow frequency band, then the power spectrum appears as in Figure 1 with only a single "spike" at that frequency. We would interpret all frequencies except this one as containing no information or at best only "noise." If, on the other hand, all frequencies contained the same amount of information or "energy" the graph of the power spectrum is as in Figure 2, a horizontal line at that energy level. This would be very much like the power spectrum of "white noise." The more usual case than either of these is the one where there are several peaks in the power spectrum graph. The percentage of energy in each band indicates the relative information content of that frequency band.

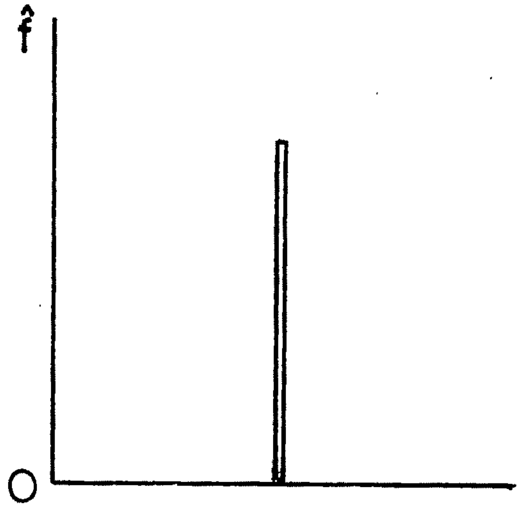


FIGURE 1. POWER SPECTRUM OF A TIME SERIES WITH ALL ENERGY CONCENTRATED IN ONE FREQUENCY COMPONENT

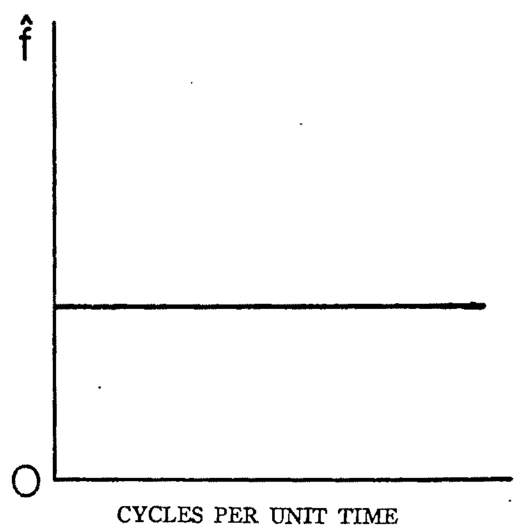


FIGURE 2. POWER SPECTRUM OF WHITE NOISE

quencies higher than the monthly frequency of .25 cycles per week, we believe that our results are not affected by the folding of the energy remaining in the frequencies above the Nyquist. Thus we suggest that we have no problem of a spurious monthly frequency from this source.

Table 1 below presents the frequency bands in terms of the length of time for a complete cycle for each series.[9] Those frequency bands for which the total energy was less than 5 per cent are not given.

TABLE 1—PERCENTAGE OF SPECTRAL POWER IN FREQUENCY BANDS[a]

Frequency	G	R	R*	FR	FR*
2.2 weeks	—	11.0	7.1	8.2	—
2.7 weeks	—	5.8	—	—	—
3.3 weeks	—	5.0	—	—	—
4.3 weeks	46.8	5.6	42.3	23.8	46.3
13 weeks	—	5.3	—	—	—

[a] The frequency indicated is the mid-point of the actual frequency band.

For the series R^* and FR^*, almost one-half of the total information in each series occurs in the monthly frequency band. Clearly these series are dominated by a monthly cycle. For R^* the only other frequency band with as much as 5 per cent of the energy is the two-week cycle. There is no other frequency band in FR^* with as much as 5 per cent of the energy.

Figure 3 presents a comparison of the power spectrum of R with that for R^*. There the extreme flattening of the entire power spectrum is evident. The monthly energy peak is reduced to contain only 5.6 per cent of the total energy of the series, striking evidence that Federal Reserve *OMO* reduce this dominant component of the net series. In some of our earlier estimates it appeared that the percentage of total energy in the business cycle frequencies (36–42 months) was increased. This does not appear in our most recent studies. We do find, however, that the small peak at two weeks is actually increased by *OMO*. By way of explanation, we are inclined to attribute this augmentation (and indeed the entire two-week cycle peak) to "folding" of the energy in frequencies above the "Nyquist." Thus, we infer that the net result of Federal Reserve *OMO* is to dampen the dominant monthly cycle in changes in member bank reserves. However, when one examines a graph of the entire power spectrum of each series (which we have not presented here), R is much more nearly flat than R^*, thus suggesting that it is more nearly a "white noise" series than is R*.

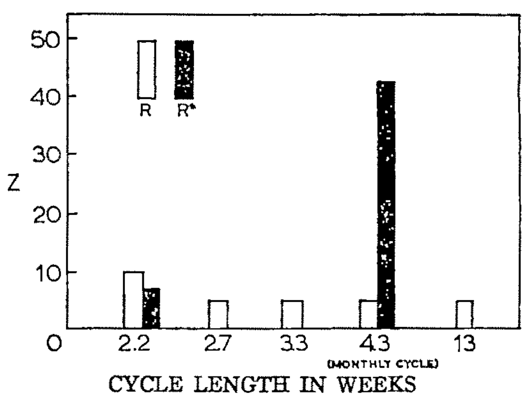


FIGURE 3. COMPARISON OF POWER SPECTRA OF R AND R^* AT FREQUENCIES CONTAINING MORE THAN 5 PER CENT OF THE TOTAL VARIANCE

The Fed is much less successful in reducing the monthly peak in FR^* than in R^*. Even here the percentage of total energy is reduced to half its former level (see Table 1 and Figure 4). Since the power in the two-week cycle frequency for FR^* is just over 4 per cent, the increase in this short cycle is smaller by a fraction than that induced by *OMO* in R. Our in-

[9] The spectral estimates were produced through the use of the FORTRAN IV procedure SPECTA written by Schotta for use on the IBM 7044 at the Davis Campus and the IBM System 360/50 at the Santa Barbara campus. The procedure is a major revision of the programs for spectral analysis contained in H. Karreman [13]. The data were not prewhitened. All estimates which we report were produced with the use of the Parzen spectral window, see E. Parzen [15, pp. 132-61], although we also estimated the spectra of all data sets using the Tukey-Hanning spectral window, see R. B. Blackman and J. W. Tukey [2], in order to check for negative estimates at low frequencies. Under the Z transform we use, the variance decomposition patterns for the two spectral windows were quite similar.

terpretation of the two-week cycle in these series is the same as for R and R^*.

The differential reduction in energy between the pairs R and R^* and FR and FR^* suggests two alternate interpretations. First, one could attribute the differential reduction to the fact that reserves are actually the Fed's policy variable with free reserves a secondary target. The second interpretation, which seems more reasonable to us, both a priori and in the light of some cross-spectral estimates presented below, is that, for the banking system as a whole, OMO are directly associated with a change in R while there is an element of discretion on the part of banks in changing FR. If banks do not immediately adjust their loan portfolios in response to OMO, then we should expect this differential. We are inclined to the view that it is the lower measure of control possessed by the Fed over FR, because of lags in portfolio adjustments of banks, which largely accounts for the difference.

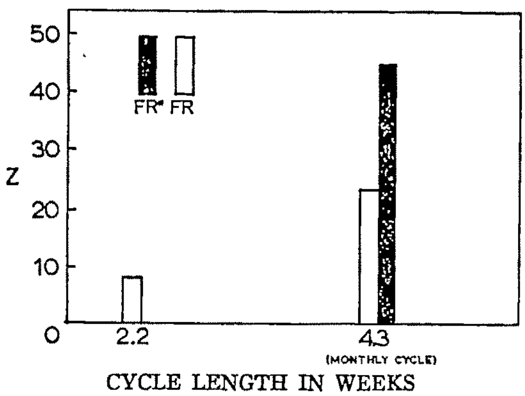


FIGURE 4. COMPARISON OF POWER SPECTRA OF FR AND FR^* AT FREQUENCIES CONTAINING MORE THAN 5 PER CENT OF THE TOTAL VARIANCE

This interpretation is reinforced by the cross-spectral statistics presented in Table 2. Cross-spectral methods relate the frequencies of one series to the corresponding frequencies of another series and yield a set of summary statistics for the frequency band which are roughly analogous to those obtained from correlation and regression. For a pair of time series samples $\{x_t, t = 1, 2, \ldots, n\}$ and $\{y_t, t = 1, 2, \ldots, n\}$, each to be estimated with m lags, the cross-spectral estimators can be obtained by first computing the pairs of cross-covariance functions for lags $i = 1, 2, \ldots, m + 1$

$$(10a)\qquad C_{xy}(i) = \frac{1}{n}\left[\sum_{t=1}^{n-(i-1)} x_t y_{t-(i-1)} - \frac{1}{n}\left(\sum_{t=1}^{n} y_t\right)\left(\sum_{t=1}^{n} x_t\right)\right]$$

$$(10b)\qquad C_{yx}(i) = \frac{1}{n}\left[\sum_{t=1}^{n-(i-1)} y_t x_{t-(i-1)} - \frac{1}{n}\left(\sum_{t=1}^{n} y_t\right)\left(\sum_{t=1}^{n} x_t\right)\right].$$

These cross-covariance functions are then weighted as in (6a) and (6b) to form the weighted cross-covariance functions,

$$(11a)\qquad C'_{xy}(i) = w(i)C_{xy}(i)$$

$$(11b)\qquad C'_{yx}(i) = w(i)C_{yx}(i)$$

where

$$i = 1, 2, \cdots, m + 1.$$

These are then transformed into the cross-spectral estimates of the co-spectrum (the "real" part of the cross-spec-

TABLE 2—CROSS-SPECTRAL STATISTICS AT THE MONTHLY CYCLE FREQUENCY G AND R^*; G AND FR^*

Series	Coherence	Gain	Phase	Γ
R^*	.86[a]	1.06[b]	−1.2[c]	−.8[c] weeks
FR^*	.81[a]	1.06[b]	−1.3[c]	−.9[c] weeks

[a] Significantly different from zero at the .95 probability level.
[b] Not different from unity at the .95 probability level.
[c] Not different from zero at the .95 probability level.

trum) and the quadrature spectrum (the "imaginary" part of the cross-spectrum). Sometimes these are referred to as the "in-phase" and "out-of-phase" components at each frequency.

$$c(i) = \sum_{j=1}^{m+1} \frac{1}{2} (C'_{xy}(j) + C'_{yx}(j)) \cos \frac{(j-1)(i-1)\Pi}{m} \tag{12}$$

$$q(i) = \sum_{j=1}^{m+1} \frac{1}{2} (C'_{xy}(j) - C'_{yx}(j)) \sin \frac{(j-1)(i-1)\Pi}{m} \tag{13}$$

for $i = 1, 2, \ldots, m + 1$ and $j = 1, 2, \ldots, m + 1$.

Just as the power spectrum for a series provides a variance decomposition in the frequency domain, the cross-spectrum provides a covariance decomposition in the frequency domain for two series. From $c(i)$ and $q(i)$ the cross-spectral statistic of coherence is calculated for the $m + 1$ lags according to

$$C^2(i) = \frac{c^2(i) + q^2(i)}{[\hat{f}_x(i)][\hat{f}_y(i)]} \tag{14}$$

for

$$i = 1, 2, \cdots, m + 1.$$

This statistic is the rough analogue of the coefficient of determination and lies in the range $0 \leq C^2(i) \leq 1$. Two series that are totally unalike are said to be ***incoherent*** while a coherence of 1.0 implies that, at the particular phase angle used, the series are not different. This statistic is a normalized cross-spectral density function.

Gain is the rough analogue of the regression coefficient of $\{x_t\}$ on $\{y_t\}$ and measures the amplification of the base series which produces the values in the series crossed on the base series.

$$G(i) = \frac{[c^2(i) + q^2(i)]^{1/2}}{\hat{f}_x(i)} \tag{15}$$

for

$$i = 1, 2, \cdots, m + 1$$

There is no precise equivalent in bivariate time series analysis in the time domain of the remaining cross-spectral statistics of phase and Γ.

It should be noted that, when we are crossing two series at some specific frequency, the only possible differences between them are in amplitude and in phase, since the frequency of oscillation is the same. Coherence provides a general measure of this similarity while gain measures the amplitude difference.

Phase is the angular measure of the shift on the time axis of the crossed (or "output") series relative to the base (or "input") series at which coherence is maximized at that frequency. The phase statistic is calculated as

$$\phi(i) = \tan^{-1}\left(\frac{q(i)}{c(i)}\right) \tag{16}$$

for

$$i = 1, 2, \cdots, m + 1.$$

We calculate a statistic, $\Gamma(i)$, which converts the phase statistic into a measure of the difference between the two series on the time axis at each frequency.

$$\Gamma(i) = \phi(i)/2\Pi\omega(j) \tag{17}$$

for

$$i = 1, 2, \cdots, m + 1.$$
$$j = 0, 1, \cdots, m.$$

and

$$\omega(j) = \frac{j\Pi}{m} \tag{18}$$

$$j = 0, 1, \cdots, m.$$

which is the measure of the frequency in terms of cycles per unit time.

As a matter of interpretation, when $\Gamma(i)$ is positive the phase shift of the series relative to the base series has been such that the base series leads the crossed series.[10]

In general it is considered profitable only to consider cross-spectral statistics for those frequencies where the series in question exhibit considerable energy. In the case of the series in Table 2 the statistics refer only to the monthly frequency. The cross-spectral statistics for G on R and G on FR are not included since the coherence generally throughout the frequencies was around .03 for each and did not differ from zero at the .95 probability level.[11]

As might be expected, if indeed *OMO* are directed toward offsetting the variance of R^* and FR^* which is contained in the monthly cycle frequency band, the coherence between R^* or FR^* and G as presented in Table 2 is large and significant. While there is some difference in amplitude of these series this, taken together with the $C^2 < 1$, merely indicates that not all of the *OMO* were defensive in nature.[12]

The substantially lower C^2 when R and FR are crossed respectively with G than is the case when R^* and FR^* are crossed

[10] For a discussion of this statistic see L. E. Borgman [5, pp. 7–9—7–20]. In order to enable the reader to interpret our $\Gamma(i)$ statistic relative to that reported in other studies we present some definitional formulae. Bendat and Piersol [1, pp. 198–99, 290–92], interpret the cross-spectral density function estimates against the background of the computation of the autocovariance functions for the time series in question according to

$$C_r = \frac{1}{N-r}\sum_{n=1}^{N-r} x_n x_{n+r} \qquad r = 0, 1, 2, \cdots, m$$

where r is the lag number, m is the maximum lag, and C_r is the estimate of the true value of the autocovariance at lag r for a time series sample $\{x_n, n = 1, 2, \ldots, N\}$. When the autocovariance function is calculated in this way the cross-spectral statistics of phase and $\Gamma(i)$ are interpreted such that $\Gamma(i) < 0$ implies that the "input" series (or base series) leads the "output" series (or crossed series) while the condition that $\Gamma(i) > 0$ implies the reverse. Since we use equation (4) above to calculate the autocovariance functions our interpretation of the sign of $\Gamma(i)$ and phase are exactly the reverse of the above for our results. We should note here that in accordance with current practice, see Fishman [9, pp. 119–21], we use equation (4) to secure the autocovariances even though, relative to the formulation above, it is a biased estimator of C_r.

[11] It should be noted here that when a random variable such as R^* is formed as in equations (1) and (2) above, then it is the case that

$$\text{var } R^* = \text{var } R + \text{var } G - 2 \text{ cov } (R, G).$$

However, our finding of a coherence throughout the frequency bands of essentially zero between R, G and between FR, G means that the last term above must drop out. Thus var R^* must be primarily explained by the sum of var R + var G. Since the objectives of *OMO* are to compensate for the var R^*, one would expect, a priori, if *OMO* are mostly defensive and are successful over time, that cov $(R, G) = 0$ and var G dominate var R^*, and further that cov (R^*, G) be large (and negative). In fact, J. H. Wood [18] provides evidence on the desired negative correlation between R^* and G. Our general findings support the large cov (R^*, G). Hence, we think our comparison of the patterns of variance decomposition of R relative to R^* out over the frequency spectrum strongly supports the hypothesis that the Fed is quite successfully engaging in defensive *OMO*. What we show is that if R is treated as a signal produced by "mixing" G with R^*, with G influenced by the factors in the process underlying the production of R^*, the effect of G is to "spread" the variance appearing in the monthly frequencies out over the remaining frequencies so that, one can say that the amount of variation of R which exists in the dominant frequency is damped. The confidence intervals for tests of H_0: $C^2(i) = 0$, as well as all tests of hypotheses below, were all taken from Bonomo and Schotta [3].

[12] This statement must be further amplified. If the pair G and R^* are analyzed in the frequency domain and are found to have differences in amplitude and phase, then this could arise because not all G were defensive in nature and/or *OMO* were not perfect. All we are prepared to advance as interpretation of the coherence findings in Table 2 is that *OMO* were not complete offsets to the factors producing R^*. We are currently analyzing the data in subperiods to see if this coherence statistic changes among postwar cycles. However, only if all *OMO* are defensive and perfectly effective will $C^2_{GR^*} = 1$. For some evidence on the percentage of *OMO* which is devoted to defensive operations see Wood [18, pp. 28–29]. Using an approach which involves regression methods applied to a time series he secures an estimate of the magnitude of *OMO* devoted to "defensive" operations which is consistent with our interpretation of the C^2 in Table 2.

respectively with G reinforces the interpretation presented earlier in the paper that the effect of *OMO* was to redistribute the variance of both series among the other frequency elements of both the series on reserves and on free reserves.

Since by equation (8) we know that the auto-covariances and cross-covariances at the zero lag are estimates of var x_t and cov x_t, y_t respectively, it can be seen from Table 3 below that var R is approximately 57 per cent of var R^* while var FR is approximately 40 per cent of var FR^*.

TABLE 3
COVARIANCE ESTIMATES AT LAG ZERO[a]

	G	*R*	*R**	*FR*	*FR**
G	66919	13023	−53896	4999	−61920
R		56986	43962	21041	8021
*R**			97860	16045	69939
FR				36513	31436
*FR**					93436

[a] Cov $(R.R) = \sigma_{R^*}^2$

This is striking evidence that, in addition to altering radically the distribution of the variances in the frequency domain, the introduction of *OMO* significantly reduces the variance of reserves and free reserves from the levels which would have prevailed in the absence of Fed action.

IV. *Summary*

To summarize our results, we have found through calculations of the power spectrum for the series R, R^*, FR, FR^*, and G that the effect of Federal Reserve open market operations since the Accord has been practically to eliminate the very strong monthly element cycling in member bank reserves, to dampen appreciably the equally strong monthly cycle in free reserves, and to reduce significantly the total variance of both series.[13]

We attribute the differential reduction in the monthly cycle between R and FR to the lower degree of direct control over free reserves on the part of the Fed and not to a primary emphasis on reserve levels as a target variable.

After we had secured our statistical results, several questions suggested themselves. First, why should there be such a pronounced monthly cycle in reserves and free reserves? We rejected the idea that this was a spurious element introduced by the subtraction of G since we could have secured R^* and FR^* by adding up the factors other than *OMO* affecting reserves.[14] In connection with the earlier study referred to on the hypothesis of independence between changes in the currency-money ratio and G, we had occasion to calculate the power spectrum for bi-weekly changes in currency outside all banks, C. We found that almost 60 per

[13] We are currently examining data for these variables for the entire period 1931–68, for individual business cycles during this period, and for certain selected sub-periods to ascertain the variance reduction and variance redistribution effects of *OMO*.

[14] One commentator has suggested that our finding of a monthly cycle peak is spurious because we use week-to-week changes in our data rather than levels. Our hypotheses concern the relationship between changes in the Federal Reserves' securities portfolio (*OMO*) and changes in variables measuring the various reserve accounts—not the relationship between levels of these variables. Thus, we have used first-differencing to produce data sets more in accord with our a priori hypotheses, not for statistical reasons. However, there are sound reasons for using a first-difference digital transform when interested in the higher frequencies. Had the necessity for eliminating specific frequency elements been our primary consideration, we probably would have used a specific high-pass filter to eliminate the strong-trending mean in the series for the variables on levels. For an evaluation of the use of a first-difference digital transform, including the effect on the power spectral density function, see Jenkins and Watts [12, pp. 205–300] and E. Malinvaud [14, pp. 413–19]. C. W. J. Granger and H. J. B. Rees [11, p. 69] also use this transform for economic time series. For an application of spectral analysis which seems to suggest the possibility of spurious peaks in the power spectrum as a result of the use of this transform, see T. J. Sargent [17, p. 165]. On the basis of the statistical sources cited above and other evidence, primarily the existence of the monthly cycle in the original series, we reject the suggestion that our finding of a monthly cycle in the series is spurious.

cent of the energy of this series was contained in the monthly cycle band. Converting G to bi-weekly changes, we continued to find a peak at the monthly cycle. In fact nearly 50 per cent of the energy of this series was then compressed into a rather narrow band around this frequency. We calculated cross-spectral statistics for this frequency which are presented in Table 4.

TABLE 4—CROSS-SPECTRAL STATISTICS AT THE MONTHLY CYCLE FREQUENCY G AND C, G AND R^*

Series	Coherence	Gain	Phase	Γ
C	.60[c]	.70[b]	.10[a]	.01 weeks[a]
R^*	.84[c]	.99[b]	.05[a]	.005 weeks[a]

[a] Not different from zero at the .95 probability level.
[b] Not different from one at the .95 probability level.
[c] Different from zero at the .95 probability level.

In terms of time units, the open market operations at this frequency and the change in currency occurred with no time lag. This is also the case between R^* and G.

When we calculate the cross-spectrum between R^* (the base series) and C, we found that for the monthly frequency, which was again the only frequency common to both series that contained a substantial amount of information, coherence was .87, gain was .75, and Γ was not different from zero.

These findings are suggestive of an interpretation of open market operations as being undertaken, or at least having the effect of existing primarily, for the purpose of offsetting a monthly cycle in changes in currency outside all banks.

A common-sense explanation of this, which is consistent with one of the original purposes of the Federal Reserve System to provide elasticity in the volume of currency for the purposes of trade, is that institutional payments arrangements are monthly (or bi-weekly) and, concurrent with the rise in the volume of check transactions around these points, there is the very common monthly practice of the "check for cash." We suspect that there would appear a very strong two-week cycle in changes in currency outside all banks if we had access to accurate daily data on this series.

Future directions for our research appear to lie in the investigation of the reasons for the monthly currency cycle. We also intend to investigate whether or not there exists a fixed relationship between the amount of float and bank debits. We have seen enough evidence to tentatively formulate the hypothesis that, in their defensive operations, the Federal Reserve is largely offsetting a rather regular and predictable monthly cycle in currency outside all banks and in float with its alteration of reserves. It may very well be this factor in open market operations which has allowed comparisons between annual changes in the money supply as a result of discretionary monetary policy and the various fixed rule policies to be so favorable to the latter.[15]

REFERENCES

1. N. S. BENDAT AND A. G. PIERSOL, *Measurement and Analysis of Random Data.* New York 1966.
2. R. B. BLACKMAN AND J. W. TUKEY, *The*

[15] In particular, Martin Bronfenbrenner [6] utilized annual data which would, if our interpretation is correct, preclude the short-run evidence we have from appearing in his results. Thus his study, because of the extremely long digital interval, was net of a major reason for open market operations. His later quarterly revision of this study [7] is still subject to the same criticism. In fact we have seen no studies of the effect on major economic variables of alternative monetary rules which utilize a data set with a digital interval short enough to analyze the phenomenon we have found to be an important aspect of Federal Open Market Operations.

Nothing in our analysis should be interpreted as suggesting that the magnitude of the Fed's monthly offset against the cyclical pattern in currency and coin outside all banks is the same; we merely show that the Fed has been engaging in this monthly offset policy.

Measurement of Power Spectra. New York 1958.

3. V. Bonomo and C. Schotta, "Statistical Tables for Time Series Analysis," Technical Paper No. 69-1, Department of Economics, Virginia Polytechnic Institute, Blacksburg 1969.
4. ———, "The Relation between Federal Open Market Operations and Variations in the Money Multiplier: A Cross-Spectral Analysis." Paper presented at the Western Economic Association Meetings, August 25, 1967, Boulder, Colorado.
5. L. E. Borgman, *Statistical Time Series.* Berkeley 1966.
6. M. Bronfenbrenner, "Statistical Tests of Rival Monetary Rules," *Jour. Pol. Econ.*, Feb. 1961, *69*, 1–14.
7. ———, "Statistical Tests of Rival Monetary Rules: Quarterly Data Supplement," *Jour. Pol. Econ.*, Dec. 1961, *69*, 621–25.
8. L. Brown, *Methods for the Analysis of Non-Stationary Time Series with Applications to Oceanography*, Technical Report HEL-16-3, Hydraulic Engineering Laboratory, College of Engineering, University of California. Berkeley 1967.
9. G. S. Fishman, *Spectral Methods in Econometrics*, Rand Report R-453-PR. Santa Monica 1968.
10. C. W. J. Granger and M. Hatanaka, *Spectral Analysis of Economic Time Series.* Princeton 1964.
11. C. W. J. Granger and H. J. B. Rees, "Spectral Analysis of the Term Structure of Interest Rates," *Rev. Econ. Stud.*, Jan. 1968, *35*, 67–76.
12. G. M. Jenkins and D. G. Watts, *Spectral Analysis and its Applications.* San Francisco 1968.
13. H. F. Karreman, *Computer Programs for Spectral Analysis of Economic Time Series*, Princeton University Econometric Research Program, Research Memo. No. 59, Princeton 1963.
14. E. Malinvaud, *Statistical Methods of Econometrics.* Chicago 1966.
15. E. Parzen, *Time Series Analysis Papers.* San Francisco 1967.
16. M. B. Priestly, "Evolutionary Spectra and Non-Stationary Processes," *Jour. Royal Stat. Soc.*, Ser. B, 1965, *27*, 204–37.
17. T. J. Sargent, "Interest Rates in the Nineteen-Fifties," *Rev. Econ. Stat.*, May 1968, *50*, 164–72.
18. J. H. Wood, *A Model of Federal Reserve Behavior*, Staff Economic Studies, No. 17, Federal Reserve Board of Governors. Washington 1966.

Reserve Policies of Central Banks and Their Implications for U.S. Balance of Payments Policy

By HELMUT A. HAGEMANN*

Under the gold-exchange standard, national currencies have supplemented gold as an international means of payment. The extent to which the fiduciary component of international reserves can be expanded depends on the reserve policies of central banks. Investigation of these policies is essential in assessing the adequacy of international reserves and the urgency of reducing the U.S. balance of payments deficit. Therefore, in this paper, we extend and test Kenen's model on reserve policies of central banks [4] and show its implications for U.S. balance of payments policy.[1]

I. *A Theory of Reserve Policies of Central Banks under the Gold-Exchange Standard*

In Kenen's model the aggregate dollar-reserve ratio of central banks, D/R, is determined by the ratio of the U.S. gold stock to her short-term liabilities to central banks, G_A/D, and the interest rate on these liabilities, r.[2]

$$D/R = f(G_A/D, r) \tag{1}$$

However, central banks' evaluation of the risks of holding dollars will not only depend on the G_A/D ratio, but on other foreign assets and liabilities of the United States as well, especially on those that are highly liquid and thus could easily affect the G_A/D ratio. Therefore, we have extended Kenen's model to include U.S. short-term assets and liabilities, SA, SL, and U.S. long-term assets and liabilities, LA, LL.

Kenen built his model around a global function for central banks' reserve-asset preferences. Our model includes functions for the portfolio policies of individual central banks and thus is able to take into account all those factors $X_i^1, \cdots, X_i^m$, which reflect the relations of country i to the world and to the United States in particular. Disregarding the fact that reserve policies of one country might depend on those of another, we can rewrite Kenen's behavioral function for the k countries outside the United States in the following way:

$$\begin{aligned} D/R &= \sum_{i=1}^{k} [D_i/R_i] \cdot W_i \qquad \text{where } W_i = R_i/R;\ D_i \geq 0;\ R_i \geq 0 \\ &= \sum_{i=1}^{k} f_i(G_A/D, r, SA, SL, LA, LL, X_i^1 \cdots X_i^m) \cdot W_i \end{aligned} \tag{2}$$

Now let us turn to how the specific economic position and environment of a country is likely to influence its reserve policies. Because of the close interde-

* The author was a teaching fellow, Harvard University from 1964 to 1966 and is now an associate in the Düsseldorf office of McKinsey & Co., Inc. He is indebted to Professors Hendrik S. Houthakker and Gottfried Haberler for many suggestions and comments on earlier drafts of this paper. The study was made possible by a grant from the Ford Foundation to Harvard University for studies in international economics.

[1] The implications of this analysis for international monetary policy are discussed elsewhere [2] [3].

[2] In the appendix, the D_i/R_i ratio of central banks is made dependent on the ratio of the U.S. gold stock to her short-term liabilities to foreign central *and* commercial banks [4, p. 584, eq. (A-3)].

pendence of the various national economies, reserve policies of a country can easily have favorable or unfavorable repercussions on that country, particularly if it is large. Thus, in formulating their reserve policies, countries will take into account such repercussions as may result from their economic environment and, specifically, from the following factors.

Economic Size and Wealth of a Country

A large country is not as free to convert dollars into gold as a smaller one. The wide publicity such a conversion is likely to attract may induce a large country to hold relatively more dollars than a smaller country. A large country may also have a greater interest in the stability and proper functioning of the international monetary system and of the world economy as a whole. Because of its size and its possible influence on the reserve policies of smaller countries, its behavior has more weight. On the other hand, a smaller country may act more independently and reduce its D_i/R_i ratio without great publicity and repercussions on its economy.

The effect of the wealth of a country on its D_i/R_i ratio is not clear on a priori grounds. On the one hand, one can argue that a wealthy country has much more to lose in a monetary upheaval than a poor country. It will therefore hold more dollars in order to prevent a breakdown of the system. On the other hand, the poor countries have a much greater preference for income and might for this reason hold more dollars than the rich countries.

The Flow and Stock of a Country's Foreign Assets and Liabilities

1. *The Current Account.* A country, which continuously runs a current account deficit and thus continuously loses foreign assets, or is highly dependent on capital inflows from other countries, is less likely to reduce its D_i/R_i ratio than a country with a continuous current account surplus, since such action could lead to restrictions on international capital movements and trade, or to deflationary measures in the United States and in turn in other countries. These balance of payments measures would then accelerate the drain on the country's foreign assets or endanger its supply of foreign capital.

On the other hand, a country with a continuous current account surplus need not fear the consequences of such restrictions and deflationary measures in other countries as much. A country may even consider deflation in other countries desirable, if it suffers from imported inflation; and a decline in the current account surplus may be welcome, if it does not want to save internationally nor assume ever higher foreign capital risks by accumulating large amounts of foreign assets, particularly if these take the form of low-yielding international reserves.

Thus, countries whose net foreign assets are continuously increasing tend to adopt a lower D_i/R_i ratio, *ceteris paribus*, than countries whose net foreign assets are continuously decreasing. However, the D_i/R_i ratio will not only depend on the aggregate net in- or outflow of foreign capital, but also on the direction and types of foreign capital flows, as well as on the stock of foreign capital already accumulated.

2. *Direction and Types of Foreign Capital Flows.* A change in the D_i/R_i ratio of a country will have a direct effect on the reserve position of the United States and only an indirect effect on the reserves of other countries through the U.S. balance of payments policies. This means that the primary repercussions of any change in the D_i/R_i ratio will be on a country's capital movements and trade with the United States. Therefore, a country which relies heavily on foreign aid and other capital

inflows from the United States is likely to adopt a higher D_i/R_i ratio than a country that does not depend on U.S. capital or that exports capital to the United States.

However, it will depend on the type of capital flows concerned. For instance, if a country receives capital in the form of direct foreign investments with high returns and associated with a change in control over assets, it may act very differently from a country that receives capital in the form of foreign aid or portfolio investments. The former may try to reduce direct foreign investment by converting dollars into gold, while the latter country may encourage the other types of capital inflows by adopting a higher D_i/R_i ratio.

3. *The Stock of Foreign Assets and Liabilities*. If a country as a whole has already a large stock of net foreign assets and has assumed large foreign capital risks, it will be less willing to accumulate even more foreign assets and incur additional risks than a country with low or negative net foreign assets. By holding gold instead of dollars, a country can decrease the foreign capital risks for the country as a whole,[3] but more importantly, it can put pressures on the balance of payments of the United States and in turn on other countries, and thereby reduce the accumulation of net foreign assets.

Of particular importance for a country's reserve policies is the stock of international reserves. If a country's reserves are large relative to its GNP or to total reserves outside the United States, it will tend to lower its D_i/R_i ratio in order to put pressure on other countries to eliminate their deficits or increase their surpluses. On the other hand, a country with low reserves will tend to adopt a higher D_i/R_i ratio in order to help other countries avoid restrictive balance of payments measures and thereby facilitate its own acquisition of reserves.[4]

However, besides the gross amount of reserves, other assets, credits, and liabilities may also influence the D_i/R_i ratio. For instance, if a central bank sells dollars to a commercial bank under a repurchase agreement, it may still regard these dollars as part of its own reserves and not change its gold holdings at all. However, statistically, such a transaction would result in a decrease of official dollar holdings and consequently in a reduction of the D_i/R_i ratio. Or, if a central bank enters a swap arrangement with the United States, it may increase its D_i/R_i ratio, as long as its own liabilities remain outstanding. It may regard dollars gained in this way as different from dollars earned through regular transactions. Or, if conditional reserves under the I.M.F. credit tranche are increased, the central bank may reduce its D_i/R_i ratio, since the credit tranche is paid in foreign currencies and is thus closer in nature to foreign exchange reserves than to gold. However, the I.M.F. gold tranche can be considered such a close substitute for gold that in this paper it is included in the gold holdings of a country.

If foreign assets of nonofficial holders become very large, there is always a danger that the central banks will be flooded with dollars once these assets are liquidated. Such large-scale liquidations are most likely to occur when reserve currencies are weak and when it is most awkward to convert dollars into gold. By holding less dollars when foreign assets of

[3] Since in many countries international reserves represent a considerable portion of total foreign assets, reserve policies have an important influence on the risk-income characteristics of the total portfolio of a country.

[4] However, according to Kenen, some countries may have such a great preference for income that they mainly accumulate dollars and only hold a certain amount of gold as a "war chest." Thus, the D_i/R_i ratio could fall if reserves increase [5]. However, this kind of behavior can more likely be expected from less developed countries, which will not be investigated here.

other institutions are large, a central bank can avoid this embarrassing situation. Such action on the part of a central bank may also be taken with a view to balancing the foreign exchange risks of the country as a whole, as we have seen above. If individuals and banks assume large exchange risks by holding foreign assets, a central bank may adopt a more conservative portfolio policy.

A Country's Foreign Trade with the Reserve-Currency Countries

Similarly, as a country will take into account the repercussions of its reserve policies on its foreign capital flows it will also consider the repercussions on its foreign trade. If a country exports primarily to reserve-currency countries, it is more likely to accept a low gold content of its reserves. By doing so, it can help the reserve-currency countries to avoid harsh balance of payments measures which would be most harmful to its own foreign trade.

Moreover, the closer the trade ties with the United States, the more likely a country will be to follow a devaluation of the dollar and avoid embarrassing capital losses on its balance sheet.[5]

Dynamic Features of an Adjustment Process

Central banks holding a large amount of reserves are not likely to change their D_i/R_i ratio radically from one quarter to the next. They will generally change the composition of their reserves only gradually, if any of the independent variables change. It is for this reason that an autoregressive structure has been introduced into the regression equations to follow.

However, the immediate impact of a change in reserves during the period of observation on the D_i/R_i ratio at the end of that period cannot be taken into account by an autoregressive term. Since frequent conversions of dollars to gold and back again to dollars would be very costly, it is most likely that central banks initially accumulate their additional reserves in the form in which they have earned them, i.e., generally in dollars. And, conversely, when central banks use their reserves to support their currencies, they will usually draw first on their foreign exchange holdings. Therefore, initially a rise in reserves is likely to increase the D_i/R_i ratio, and a decline in reserves is likely to decrease that ratio.

II. *Regression Analysis of the Reserve Policies of Central Banks*

The view that the portfolio policies of central banks are determined by politics regardless of any economic considerations seems to be widespread. The only major attempt to test reserve-asset preferences empirically was made by Kenen in 1963 [5]. However, his analysis is limited in scope. The only variables which enter into his regressions are the total amount of reserves of a particular country, its foreign exchange holdings, and the changes in each of these variables. No attempt was made to take into account all the other variables that also have an important influence on reserve policies—all those that determine the international liquidity position of the United States and those that characterize the position of an individual country within the entire structure of international trade and capital movements. In the following, we will investigate these complex relationships by testing the various determinants of the D_i/R_i ratio. Because in most countries dollar holdings of central banks are not published, we have investigated the foreign exchange-reserve ratio instead of the dollar-reserve ratio. We excluded all sterling-area countries from our analysis, as they have tradi-

[5] It will, of course, still have foregone a capital gain, but that can be defended much easier before the public.

tionally held a large portion of their reserves in pounds.[6]

The following nomenclature has been adopted in these analyses. All variables are in $ million, if not otherwise indicated:

- r = Average tender rate for three-month Treasury Bills, per cent [6]
- G_A = U.S. gold stock [6]
- D = U.S. short-term liabilities to official institutions [6]
- SA = U.S. short-term foreign assets excluding the U.S. gold stock [8]
- SL = U.S. short-term foreign liabilities excluding U.S. liabilities to official institutions [8]
- LA = U.S. long-term foreign assets [8]
- LL = U.S. long-term foreign liabilities [8]
- DF = U.S. balance of payments deficit, liquidity basis, seasonally adjusted [8]
- R = Total reserves outside the United States [6]
- R_i = Reserves of country i, including the I.M.F. gold tranche [6]
- D_i = Foreign exchange holdings of country i [6]
- CT_i = I.M.F. credit tranche of country i [6]
- BA_i = Foreign assets of the commercial banks in country i [6]
- BL_i = Foreign liabilities of the commercial banks in country i [6]
- DI_i = Increase in the value of U.S. direct investment in country i [8]
- PI_i = Net increase in U.S. long-term portfolio investment in country i [10]
- FA_i = Net U.S. foreign assistance to country i [9]
- CD_i = Deficit on the current account of country i [6]
- X_i = Exports of country i [6]
- XR_i = Exports to the reserve-currency countries (United States and United Kingdom) by country i [7]

Time Series Analysis for 11 Major Countries

The time series analysis is limited to 11 major countries outside the sterling area whose reserves exceed $1 billion. The reserves of these 11 countries account for approximately 75 per cent of international reserves outside the U.S. and the sterling area, and therefore have the greatest importance to U.S. balance of payments policies.

This analysis is based on quarterly data. However, for three reasons we have excluded the initial observations in the early 1950s and started with the period that showed the highest D_i/R_i ratio. First, in the early postwar years the portfolios of many central banks did not seem to be in equilibrium. At the time when there was a shortage of international reserves in general, many central banks complained more about a shortage of dollars than a shortage of gold. During the subsequent period of adjustment to the equilibrium D_i/R_i ratio, some central banks exchanged gold against dollars. Others, being more reserved about the development of the new gold-exchange standard, adjusted more slowly to a higher D_i/R_i ratio by accumulating additional dollars, only as they earned them from balance of payments surpluses. Secondly, random changes in the foreign exchange reserve ratio were often quite large, since reserves were still very low; and thirdly, changes in non-dollar exchange holdings were of relatively greater importance.

The time series analysis shows that the U.S. liquidity ratio, G_A/D, is very significant in all cases (Table 1).[7] When the

[6] At the end of 1965, these countries held 74 per cent of all British short-term liabilities to official institutions.

[7] The G_A/D ratio becomes even more significant, if we take annual data for our regressions.

TABLE 1—REGRESSIONS FOR 11 MAJOR COUNTRIES, QUARTERLY DATA

Country	Estimated Equation			T-Test[a] (Null-Hypothesis)	R^2 (per cent)	*DGF*	F-Ratio[b]	Standard Error of Estimate	Durbin Watson
Austria	$(D_i/R_i)_t=$	0.2047		3.1	97.4	42	527.0 (26.4)	0.0224	1.43
		0.0618	$(G_A/D)_t$	3.2					
		$0.2535\cdot10^{-4}$	$(SA-SL)_t$	2.8					
		0.5826	$(D_i/R_i)_{t-1}$	4.7					
Belgium-Luxembourg	$(D_i/R_i)_t=$	-0.0112		-0.8	91.0	51	129.5 (13.7)	0.0237	1.38
		0.0165	$(G_A/D)_t$	2.4					
		$0.2100\cdot10^{-3}$	ΔR_i	4.7					
		0.7445	$(D_i/R_i)_{t-1}$	9.7					
		$0.3020\cdot10^{-4}$	BL_{it}	2.0					
Canada	$(D_i/R_i)_t=$	0.2103		3.4	75.9	57	35.3 (9.2)	0.0625	1.34
		0.0317	$(G_A/D)_t$	3.7					
		$-0.2726\cdot10^{-4}$	SL	-2.5					
		$0.1519\cdot10^{-3}$	ΔR_i	3.0					
		0.3884	$(D_i/R_i)_{t-1}$	5.2					
		$0.1330\cdot10^{-3}$	BL_{it}	3.7					
France (eq. (1))	$(D_i/R_i)_t=$	1.0004		5.1	84.6	21	28.9 (14.0)	0.0310	2.04
		$0.4272\cdot10^{-4}$	$(SA-SL)_t$	3.4					
		$0.1171\cdot10^{-3}$	ΔR_i	2.1					
		0.4596	$(D_i/R_i)_{t-1}$	4.5					
		$-0.9703\cdot10^{-3}$	CT_{it}	-4.7					
France (eq. (2))	$(D_i/R_i)_t=$	0.2118		0.8	46.1	22	6.28 (26.6)	0.0567	0.81
		0.3264	$(G_A/D)_t$	2.3					
		$-0.7878\cdot10^{-4}$	SL_t	-3.4					
		$0.1795\cdot10^{-3}$	ΔR_t	2.9					
Germany	$(D_i/R_i)_t=$	0.0263		2.2	98.0	49	808.0 (26.4)	0.0244	1.50
		0.0320	$(G_A/D)_t$	4.4					
		$0.9717\cdot10^{-4}$	ΔR_t	9.9					
		0.7667	$(D_i/R_i)_{t-1}$	17.0					
Italy	$(D_i/R_i)_t=$	-0.6724		-4.2	90.1	31	70.5 (13.8)	0.0479	1.26
		0.3452	$(G_A/D)_t$	9.4					
		0.0432	r_t	2.4					
		$0.8713\cdot10^{-4}$	$(SA-SL)_t$	4.0					
		$0.1286\cdot10^{-4}$	$(LA-LL)_t$	5.5					
Japan	$(D_i/R_i)_t=$	0.3362		3.6	93.7	49	144.5 (9.2)	0.0266	1.78
		0.0131	$(G_A/D)_t$	1.8					
		$-0.3539\cdot10^{-4}$	SL_t	-3.4					
		$0.1034\cdot10^{-4}$	$(LA-LL)_t$	3.0					
		$-0.2234\cdot10^{-4}$	DF_t	-2.5					
		0.5879	$(D_i/R_i)_{t-1}$	6.1					
Netherlands	$(D_i/R_i)_t=$	0.0030		0.3	91.6	58	210.8 (26.3)	0.0401	2.04
		0.0304	$(G_A/D)_t$	4.0					
		$0.2561\cdot10^{-3}$	ΔR_t	3.8					
		0.5752	$(D_i/R_i)_{t-1}$	6.4					

[a] For the reader's convenience the T-Test for the Null-Hypothesis is given rather than the standard error. For instance, if $t=2$, the coefficient is significantly different from zero at the 0.05 level.

[b] The F-ratio for the whole equation is given and can be compared with the F-ratio which indicates a level of significance of 0.01 (in parentheses).

TABLE 1—REGRESSIONS FOR 11 MAJOR COUNTRIES, QUARTERLY DATA (Cont'd)

Country	Estimated Equation		T-Test[a] (Null-Hypothesis)	R^2 (per cent)	DGF	F-Ratio[b]	Standard Error of Estimate	Durbin Watson
Spain	$(D_i/R_i)_t = -0.0258$		-0.9	96.4	21	185.0 (26.7)	0.0266	1.87
	0.1601	$(G_A/D)_t$	5.5					
	$0.4756\cdot10^{-8}$	ΔR_t	5.0					
	0.5638	$(D_i/R_i)_{t-1}$	7.8					
Sweden (eq. (1))	$(D_i/R_i)_t = 0.0395$		2.4	96.2	51	426.0 (26.3)	0.0194	1.62
	0.0084	r_{t-1}	2.5					
	$0.5282\cdot10^{-8}$	ΔR_t	6.6					
	0.8907	$(D_i/R_i)_{t-1}$	28.7					
(eq. (2))	$(D_i/R_i)_t = 0.1901$		3.1	77.8	51	59.6 (26.3)	0.0467	0.36
	0.0461	$(G_A/D)_t$	3.6					
	$0.4840\cdot10^{-4}$	SA_t	7.4					
	$-0.4724\cdot10^{-8}$	BA_{tt}	-2.9					
Switzerland	$(D_i/R_i)_t = 0.0050$		0.7	78.7	49	60.4 (26.4)	0.0189	2.07
	0.0079	$(G_A/D)_t$	2.3					
	$0.5324\cdot10^{-4}$	ΔR_t	2.7					
	0.7464	$(D_i/R_i)_{t-1}$	8.2					

ratio of the U.S. gold stock to her short-term liabilities falls, the risk in holding dollars is increased and, consequently, central banks tend to lower their D_i/R_i ratio.

This relationship has dynamic implications. For instance, if the United States incurs a balance of payments deficit thus reducing its liquidity ratio, central banks will, *ceteris paribus*, lower their D_i/R_i ratio by exchanging dollars into gold. This will further reduce the liquidity ratio and lead to a second round of adjustments [4, pp. 579–82].

However, the adjustments to a change in any of the independent variables are not instantaneous, as the presence of the autoregressive term $(D_i/R_i)_{t-1}$ in 10 out of 13 equations shows. The particularly strong and separate influence of a change in reserves during a quarter on the D_i/R_i ratio at the end of that quarter is indicated by the presence of ΔR_i in 9 out of 13 equations. Central banks that gain reserves tend initially to accumulate foreign exchange before converting it into gold and, similarly, to reduce their foreign exchange holdings, when their reserves decline, before they adjust their gold holdings.

The interest rate on dollars is of relatively minor importance. Only in Italy and Sweden (eq. (1)) did it have any significance.

Far greater is the influence of U.S. short-term assets and liabilities (other than those to official institutions) on the reserve policies of central banks. The liabilities of the United States represent a burden on her future balance of payments and thus increase the risk to foreign dollar holdings. In Austria, Canada, France (eq. (1) and eq. (2)), Italy, and Japan, U.S. short-term liabilities played an important role in determining reserve policies of central banks. However, in Austria, France (eq. (1)), and Italy, U.S. short-term assets were regarded as a factor offsetting the risk created by her liabilities.

The overall net long-term position of the United States was only significant in the case of Italy and Japan, and the U.S. balance of payments deficit (liquidity

basis) had a significant influence only on the reserve policies of Japan. Lack of data prevented us from testing the influence of foreign assets and liabilities of the banking sector in each of the 11 countries.[8] However, it appears that, at least in Canada and Belgium, the relatively large foreign liabilities of the banking sector induced the central bank to hold more foreign exchange than it would otherwise have done, and thus avoid forcing other countries to withdraw their assets. In Sweden (eq. (2)) on the other hand, foreign assets of commercial banks seem to have made the central bank more cautious about the accumulation of foreign exchange, since the repatriation of these assets would lead to large foreign exchange holdings, which the central bank might then be unable to convert into gold.

Of course, these foreign assets and liabilities are very heterogeneous and have probably changed a great deal in composition and quality over time; in addition, they are not the only foreign assets and liabilities of a country. The holdings of nonbank institutions and individuals may also be important—indeed the size and structure of the entire foreign investment position of a country is likely to influence the way in which policy-makers evaluate the proper D_i/R_i ratio. More work is needed to test these complex relationships and to clarify the role that central banks play, or should play, in altering the size and composition of foreign assets and liabilities of a country as a whole through their reserve policies.[9]

[8] Data were available for Austria, Belgium, Canada, Germany, and Sweden. However, some of the early quarterly observations had to be derived by linear interpolation of yearly data: in Austria for the years 1954–57, in Belgium for 1952–57, in Canada for 1951 and 1952, in Germany for 1953–56, in Sweden only foreign liabilities of commercial banks had to be interpolated for the years 1950–56.

[9] Of the conditional reserves the I.M.F. credit tranche was only significant in the case of France (eq. 1). Whenever the French credit tranche increased, France reduced her D_i/R_i ratio.

Cross-Section Analysis for 18 Countries, 1964

This analysis is confined to all countries outside the sterling area with exports over $1 billion. These 18 countries have a major stake in world trade and represent a group of all major developed and semi-developed countries outside the sterling area and the United States.[10] We estimated two different equations and recomputed them for the same 18 countries excluding France (Table 2).

One of the most interesting findings of this cross-section analysis is that, in this sample, countries have responded unfavorably to U.S. direct investment in their countries. Their hidden or declared reluctance to finance direct U.S. investment by accumulating short-term dollars is reflected in their portfolio policies. If U.S. direct investment in a developed country is large in proportion to its GNP, the country tends to reduce the dollar content of its reserves, thus putting pressure on the United States to restrict these capital outflows. On the other hand, countries which receive U.S. long-term capital in other forms have adopted a higher D_i/R_i ratio than those recording a net capital outflow to the U.S. This means that the official institutions have generally welcomed U.S. long-term capital, unless it is associated with a shift of control to U.S. investors (eq. (2a) (2b)).

The D_i/R_i ratio will be even higher, if the country has continuous large deficits on the current account relative to its GNP, or if it sells a large portion of its exports to the reserve-currency countries. By accumulating a large portion of its reserves in foreign exchange and not converting it into gold, a country can help to reduce the danger that the reserve-currency countries, and in turn other countries, might restrict international capital flows and trade, or

[10] Conclusions drawn from this group of countries should not be applied to the less developed countries.

TABLE 2—CROSS-SECTION REGRESSIONS FOR 18 COUNTRIES, 1964 INCLUDING/EXCLUDING FRANCE

	Estimated Equation			T-Test (Null-Hypothesis)	R^2 (per cent)	DGF	F-Ratio[a]	Standard Error of Estimate
(1a) with France	$D_i/R_i=$	0.3555		5.1	90.8	11	18.1	0.0854
		$-0.2294\cdot 10^{-2}$	DI_i/GNP_i	−4.4			(7.8)	
		3.7744	CD_i/GNP_i^{b}	2.9				
		0.5533	XR_i/X_i	3.1				
		-0.1654	R_i/R	−2.5				
		$0.3914\cdot 10^{-5}$	GNP_i	1.9				
		$0.1178\cdot 10^{-3}$	GNP_i/Cap_i	3.3				
(1b) without France	$D_i/R_i=$	0.3488		6.6	94.6	10	29.3	0.0652
		$-0.2331\cdot 10^{-2}$	DI_i/GNP_i	−5.9			(7.9)	
		3.5670	CD_i/GNP_i^{b}	3.6				
		0.4834	XR_i/X_i	3.5				
		-0.1788	R_i/R	−3.5				
		$0.5140\cdot 10^{-5}$	GNP_i	3.2				
		$0.1319\cdot 10^{-3}$	GNP_i/Cap_i	4.8				
(2a) with France	$D_i/R_i=$	0.6311		12.2	85.9	12	14.6	0.1014
		$-0.2624\cdot 10^{3}$	DI_i/GNP^{c}	−4.3			(9.9)	
		6.1483	CD_i/GNP_i^{c}	3.8				
		$0.3716\cdot 10^{2}$	PI_i/GNP_i^{c}	3.9				
		$0.2543\cdot 10^{3}$	FA_i/GNP_i^{c}	2.3				
		-1.2159	R_i/GNP_i	−2.3				
(2b) without France	$D_i/R_i=$	0.6707		12.9	88.1	11	16.3	0.0925
		$-0.2503\cdot 10^{3}$	DI_i/GNP_i^{c}	−4.5			(10.0)	
		5.5586	CD_i/GNP_i^{c}	3.7				
		$0.3354\cdot 10^{2}$	PI_i/GNP_i^{c}	3.8				
		$0.1477\cdot 10^{2}$	FA_i/GNP_i^{c}	1.3				
		-1.4903	R_i/GNP_i	−3.0				

[a] The number in parentheses is the F-ratio for a level of significance of 0.01.
[b] 1961–64 average.
[c] 1962–64 average.

follow deflationary policies, which would widen the country's current account deficit further and be most harmful to its foreign trade.

Both R_i/GNP, and R_i/R had a negative correlation with the D_i/R_i ratio, which indicates that a country whose reserves are large relative to its GNP or to the total amount of reserves outside the United States becomes more cautious in its reserve policies and adopts a higher D_i/R_i ratio than countries with relatively low reserves. However, if a country is very large or very wealthy, it will, *ceteris paribus*, adopt a higher D_i/R_i ratio than a small or poor country, as indicated by the positive correlation of GNP_i and GNP_i per capita with the D_i/R_i ratio.[11] Thus, in this sample, the wealthy countries have adopted a higher D_i/R_i ratio than the poorer countries, because to the wealthy countries even large devaluation losses may be less harmful than a disruption of international trade and capital movements. The greater income preferences of poor countries did not play a role in this group of developed and semideveloped countries.

Analysis of the Residuals

1. The Dollar Crisis in 1960. Although the gold holdings of the 11 major countries

[11] The correlation which might be expected between GNP_i per capita on the one hand and R_i/GNP_i and R_i/R on the other is below 35 per cent.

together increased by $470 million, only France, Italy, and Spain have increased their gold stocks at the expense of their foreign exchange holdings, and only for Spain was the actual D_i/R_i ratio significantly below the computed value at the 0.05 level of significance (Table 3). In all other countries, there seem to have been no drastic changes in portfolio during the fourth quarter of 1960. Thus, despite the large acquisitions of gold, the major central banks have not departed from their portfolio policies, but neither have they helped the United States to meet the strong private demand for gold.[12]

2. Strong Demand for Gold in 1965. Although gold production continued to rise and Soviet sales amounted to a record $550 million, the world monetary gold stock increased by only $245 million or less than 0.5 per cent. This was the smallest increase since 1952. More than $1.7 billion went into private hoards and consumption, about 70 per cent more than was absorbed in 1960, the year of the dollar crisis. However, in contrast to 1960, central banks seem to have participated in a general shift towards gold in 1965.

The major reasons for the strong demand for gold by individuals and central banks was the fear of the devaluation of the pound and the possible effect it might have on the dollar.[13] Some central banks may also have expected further drawings by the Bank of England under the credit arrangements and may have reduced their foreign exchange holdings to prevent excess accumulation of foreign currencies in the future. The escalation of the Vietnam war during the latter half of 1965 may also have created additional uncertainty.

For the first time since the war, U.S. short-term liabilities to official institutions declined in a deficit year. While in 1960 the accumulation of gold in these 11 countries was accompanied by a large increase in foreign exchange holdings, the substantial acquisition of gold in 1965 coincided with an even larger decline in foreign exchange holdings.[14] This substitution took place in 10 out of 11 countries and amounted to roughly $2.4 billion (Table 3). Japan was the only country that increased its foreign exchange holdings, but still reduced its D_i/R_i ratio.

Comparison of the actual with the computed D_i/R_i ratios for all quarters of 1965 shows that in 39 out of the 52 quarters actual D_i/R_i ratios were below the expected values. Although the deviation was significant at the 0.05 level in only one instance, and no radical departure from previous portfolio policies is yet indicated, a more cautious attitude prevailed among the major central banks in 1965. They seem to be increasingly reluctant to add to their unconditional holdings of foreign exchange, even though they might still be willing or be forced by political pressures to accumulate additional foreign exchange temporarily or to extend credits to other countries in time of crisis.

3. Extrapolation of the Time Series to 1966. All our regressions are based on data that include the last quarter of 1965. Since computations were made, all data for 1966 have become available so that we were able to test the validity of our equations in explaining reserve policies of central banks in 1966.

Deviations are generally small and, with three exceptions (Switzerland, Spain, France), not significant at the 0.05 level (Table 3). Swiss foreign exchange holdings for the second and fourth quarters are sig-

[12] P. B. Kenen, [5, p. 64] reached a slightly different conclusion, as he detected some cooperation with the United States during the dollar crisis of 1960.

[13] The interdependence between the two reserve currencies with special emphasis on the effects of the sterling crisis on the willingness of other countries to accumulate dollars has been discussed in my article [1].

[14] The objection may be raised that the decline in foreign exchange holdings was primarily due to the repayment of credits by Great Britain. However, the decline of sterling liabilities to all of Western Europe was less than $100 million in 1965.

TABLE 3—RESIDUAL ANALYSIS OF THE TIME SERIES

Country	Residual Analysis for the Fourth Quarter of 1960			Residual Analysis for 1965						Extrapolation to 1966				Standard Error of Estimate
	ΔG_i $ million	ΔD_i $ million	$D_i/R_i-\widehat{D_i/R_i}$ Residual (per cent)	ΔG_i $ million	ΔD_i $ million	$D_i/R_i-\widehat{D_i/R_i}$ Residual (per cent)				$D_i/R_i-\widehat{D_i/R_i}$ Residual (per cent)				(per cent)
						I	II	III	IV	I	II	III	IV	
Austria	—	− 6	−0.7	+100	−106	−1.5	−0.9	0.0	−0.6	−2.8	2.2	3.7	4.2	2.2
Belgium-Luxembourg	+ 76	+127	2.4	+107	−103	−0.7	−4.5	− 1.0	−0.7	−1.9	−2.0	− 1.6	− 4.3	2.4
Canada	− 9	+ 23	5.9	+125	−135	−1.5	−2.7	−13.6	−9.4	−1.2	−3.4	− 5.0	−11.3	6.3
France (eq. (1))	+ 14	− 54	−2.5	+977	−623	−1.5	−2.6	− 1.3	−1.8	2.3	1.7	9.5	7.5	3.1
(eq. (2))			1.0			1.5	−5.3	− 8.0	−8.7	−7.1	−9.6	−12.4	−11.5	5.7
Germany	+ 83	+316	1.4	+162	−778	2.3	−2.5	− 0.2	−0.5	−0.4	−1.1	0.0	− 0.3	2.4
Italy	+ 17	− 21	−0.8	+297	−109	6.2	−1.3	− 2.1	0.3	−2.9	−6.0	− 0.7	− 0.8	4.8
Japan	—	+166	3.2	+ 24	+ 74	3.5	−4.7	0.3	−1.8	−3.1	−3.9	− 3.9	− 2.5	2.7
Netherlands	+ 5	+ 88	−1.9	+ 68	− 94	1.5	−1.5	0.6	1.8	0.5	−0.4	0.8	1.5	4.0
Spain	+ 80	− 12	−6.0	+194	−333	0.2	−2.7	− 2.2	2.1	0.8	−6.9	− 4.8	− 4.8	2.7
Sweden (eq. (1))	− 1	− 10	0.7	+ 13	− 61	−1.8	−2.8	− 1.9	1.3	−3.5	−1.5	− 0.9	− 0.6	1.9
(eq. (2))			1.6			−3.0	−2.9	− 6.7	−5.9	−5.9	−4.2	− 3.2	+ 1.6	4.7
Switzerland	+205	+ 5	−1.8	+317	−193	−1.6	−0.8	1.5	−2.7	2.5	4.6	− 1.1	4.6	1.9
All 11 Countries	+470	+622		+2.384	−2.461									

TABLE 4—RESIDUAL ANALYSIS OF THE CROSS-SECTIONAL REGRESSIONS FOR 1964 (PER CENT)

Country	D_i/R_i Actual	Equation (1a) $\widehat{D_i/R_i}$ Computed	$D_i/R_i-\widehat{D_i/R_i}$ Residual (a)	Standard Error of Estimate	Equation (2a) $\widehat{D_i/R_i}$ Computed	$D_i/R_i-\widehat{D_i/R_i}$ Residual (a)	Standard Error of Estimate
				8.5			10.1
Austria	49.0	46.0	2.9		42.3	6.7	
Belgium-Luxembourg	24.6	34.4	−9.7	−x	36.3	−11.6	−x
Denmark	80.5	77.3	3.2	+x	73.8	6.7	
Finland	74.0	74.8	−0.8		71.3	2.7	
France	24.0	39.9	−15.8	−xx	38.3	−14.3	−x
Germany	34.5	23.8	10.8	+x	37.0	−2.5	
Italy	41.1	33.9	7.2		48.2	−7.1	
Netherlands	16.9	26.1	−9.3	−x	20.9	−4.0	
Norway	85.5	81.7	3.8		86.8	−1.3	
Spain	52.1	50.6	1.5		58.3	−6.1	
Sweden	71.4	67.3	4.0		51.8	19.6	+x
Switzerland	12.7	11.1	1.7		6.9	5.9	
Canada	57.6	58.9	−1.3		60.9	−3.3	
Argentina	53.6	47.9	5.7		46.2	7.4	
Brazil	75.0	84.2	−9.2	−x	86.2	−11.2	−x
Mexico	63.3	65.3	−1.9		61.9	1.4	
Venezuela	47.3	43.7	3.6		44.3	2.9	
Japan	74.1	70.5	3.6		65.9	8.1	

x = Deviation significant at 0.32.
xx = Deviation significant at 0.05.
(a) = Discrepancies due to rounding.

nificantly larger than expected, possibly because of swap arrangements in connection with the support of the pound. Deviations for the third and fourth quarters in France are in different directions for the two equations. In eq. (1) the increase in the I.M.F. credit tranche has pushed the estimated D_i/R_i ratio below zero, indicating that France would hold foreign exchange only for transactions purposes. According to eq. (2), the French D_i/R_i ratio should still have been around 20 per cent.

In conclusion, we can say that the estimated "portfolio selection curves" of these countries have remained stable and continued to explain their reserve policies in 1966.

4. Residual Analysis of the Cross-Section Regressions. The residual analysis of the cross-section regression for 1964 shows that, in the two equations, positive deviations of more than the standard error were registered for Denmark, Germany, Norway, and Sweden and negative deviations for Belgium and France. Only in France was the deviation in the first equation significant at the 0.05 level (Table 4). By excluding France, the standard error could be reduced considerably and a better fit could be achieved for both equations (Table 2, eq. (1b) (2b)).

The extent to which the French reserve policy differs from the policies in other developed countries becomes even more apparent if we compute the French D_i/R_i ratio on the basis of the cross-section regressions for the remaining 17 countries. Judging from the behavior of other countries, one would expect France to hold about half of her reserves in foreign exchange. However, in view of the strained relationship France has with the United

States, it should not be too surprising if noneconomic considerations have played a considerable role in determining her reserve policies.

III. *Implications of our Analysis for U.S. Balance of Payments Policies*

Before we draw any policy conclusions on the basis of our analysis, we wish to emphasize the limitations of this procedure. We realize that past reserve policies may be revised drastically in the face of changed circumstances. Our equations are based on data from a period when the breakdown of the entire international monetary system was a remote possibility. Now, as this danger is growing, central banks may depart from past policies. Some may increase their D_i/R_i ratios to help forestall such a crisis, others may reduce their D_i/R_i ratios below the level we would expect from our equations to reduce the risk of suffering large capital losses. It is with these reservations in mind that we have drawn the following conclusions from our equations.

The Result of Continuing the U.S. Balance of Payments Deficit

The ability of the United States to finance balance of payments deficits depends on three crucial factors: (a) the size of the monetary gold stock; (b) the willingness of central banks to accumulate dollars as international reserves vis-à-vis gold; and (c) the relative distribution of international reserves among countries outside the United States.

While the monetary gold stock has started to decline and reserves are shifting to European countries, whose reserves have a relatively high gold content, the ability of the United States to finance further deficits will depend even more on the willingness of central banks to accumulate dollars.

The dynamic nature of the reserve policies of central banks and their dependence on the U.S. liquidity ratio, G_A/D, implies great uncertainty regarding changes in the U.S. gold stock. As we have seen, the impact of a U.S. balance of payments deficit on the U.S. gold stock is not felt immediately. Central banks only gradually adjust the composition of their reserves and convert dollars into gold. This leads to a further reduction in the liquidity ratio and thus to another round of adjustments.

If we assume that the U.S. incurs a cumulative balance of payments deficit (defined as $-\Delta G_A+\Delta D$) of $5 billion over the period 1966–1969, which is approximately equal to the deficit she ran over the previous four years,[15] what will be the gold outflow from the U.S. after all adjustments have taken place? In order to assess the magnitude of the gold drain, we will make the following assumptions: (a) All foreign exchange holdings of the 11 major countries are in dollars; (b) The relative distribution of reserves among these 11 countries remains the same as at the end of 1965; (c) The total amount and the composition of reserves of all countries other than the 11 countries remain at their 1965 level; and (d) The monetary gold stock and all variables other than G_A, D, D_i, R_i remain at their 1965 level.

Now we are able to derive the aggregate "portfolio selection curve" for the 11 major countries from Table 1:

$$D^*/R^* = \sum_{i=1}^{11} [D_i/R_i]\cdot W_i \tag{3}$$

$$\text{where } W_i = R_i/R^*;\ D_i \geq 0;\ R_i \geq 0$$

Since in equilibrium $\Delta R_i=0$, $(D_i/R_i)_{t-1}$

[15] A U.S. balance of payments deficit is defined here as the decrease in the U.S. gold stock plus the increase in U.S. short-term liabilities to official institutions. According to this definition, the cumulative deficit of the United States during 1962–1965 was equal to $4.8 billion.

$=(D_i/R_i)_t$, $DF_t=0$, and $r_{t-1}=r_t$, the equilibrium form of equation (3) for the 11 countries becomes:

$$(4)\quad D^*/R^* = 0.3796 + 0.1114(G_A/D) + 0.2757\cdot 10^{-4}SA - 0.3670\cdot 10^{-4}SL + 0.0316\cdot 10^{-4}(LA - LL) + 0.0076r + 0.0778\cdot 10^{-4}BL_{\text{Belg.}} + 0.1879\cdot 10^{-4}BL_{\text{Can.}} - 3.2515\cdot 10^{-4}CT_{\text{Fr.}}$$

If X is the gold outflow after 1965 and BD is the U.S. balance of payments deficit, reserves of the 11 countries will increase by BD and official dollar holdings by BD minus X. The equilibrium equation thus becomes:

$$(5)\quad \frac{D^*_{65} + BD - X}{R^*_{65} + BD} = 0.3796 + 0.1114\frac{G_{A65} - X}{D_{65} + BD - X} + 0.2757\cdot 10^{-4}SA - 0.3670\cdot 10^{-4}SL + 0.0316\cdot 10^{-4}(LA - LL) + 0.0076r + 0.0778\cdot 10^{-4}BL_{\text{Belg.}} + 0.1879\cdot 10^{-4}BL_{\text{Can.}} - 3.2515\cdot 10^{-4}CT_{\text{Fr.}}$$

$$D_i \geq 0 \qquad \text{for all 11 countries}$$

Solving for X, we get

$$X = \$5{,}520 \text{ million}$$

If central banks followed portfolio policies according to the estimated equations, a U.S. deficit of $5 billion would lead, after all adjustments, to a gold outflow from the United States of $5.5 billion under the above assumptions, reducing the U.S. gold stock to about $8.5 billion.

Thus it will become even more critical for the U.S. to correct her external imbalance, as past deficits will continue to cause gold outflows, and future deficits are likely to result in gold outflows even larger than the deficits.

The Effect of U.S. Balance of Payments Measures on the Reserve Policies of Central Banks and Their Repercussions on the U.S. Gold Stock

Although the advantages and disadvantages of various balance of payments measures have been widely discussed, mention is generally not made of the effect these measures have on the willingness of other countries to accumulate dollars as international reserves.[16] Decisions to change the composition of international reserves can have a serious impact on the U.S. gold stock, outweighing the beneficial effect the balance of payments measures may have had.

From equation (4), we can derive the marginal effects on the U.S. gold stock for a change in any of the variables that may be affected by U.S. balance of payments policies under the assumptions made above. Differentiating G_A with respect to these variables and substituting the equilibrium values for 1965, $G^E_{A65}=\$13{,}934$ million and $D^E_{65}=\$12{,}906$ million, we obtain the change in the U.S. gold stock (after all adjustments) resulting from a change in any of those variables at the end of 1965:

$$\frac{\partial G_A}{\partial SA} = 0.9429;$$

$$\frac{\partial G_A}{\partial SL} = -1.2551;$$

$$\frac{\partial G_A}{\partial(LA - LL)} = 0.1081;$$

$$\frac{\partial G_A}{\partial r} = \$260 \text{ million}$$

[16] Devaluation is the only balance of payments measure where this effect is considered.

The marginal effects of SA and SL on G_A are surprisingly high, but a large part of this effect is due to the adjustments and will not be felt immediately. If, for example, the U.S. repatriated short-term assets, the decrease in short-term assets would induce some central banks to convert dollars into gold. The resulting reduction in the G_A/D ratio will in turn induce other central banks to adjust their portfolios. After all these adjustments and the adjustments caused by the autoregressive terms, the repatriation of $1 billion of short-term assets would finally cause a gold outflow of about $0.9 billion. The marginal effect of a change in short-term liabilities on the U.S. gold stock is even larger.

Though these figures must be taken with extreme caution, they indicate that an improvement in the balance of payments attained by repatriating short-term assets or by increasing short-term liabilities may not stop the drain on the U.S. gold stock. In assessing the risks involved in accumulating dollars as international reserves, central bankers seem to be very sensitive to any deterioration in the U.S. short-term foreign investment position (official and private), whatever kinds of short-term assets or liabilities may be affected. On the other hand, the long-term investment position of the United States has comparatively little effect on the willingness of central banks to hold dollars. Therefore, restrictions on long-term capital movements are likely to be more successful in reducing the drain on the U.S. gold stock than restrictions on short-term capital flows. But as the cross-section regressions show, it depends on the type of capital movement affected. A cut in foreign aid or pressures to prepay debts will tend to reduce the D_i/R_i ratios, while the curtailment of direct investment in the developed countries may not only lead to an improvement in the balance of payments, but might also induce the developed countries to hold more dollars. From this point of view, the Interest Equalization Tax was probably much less beneficial than the limitation of direct investments.

Balance of payments measures affect different countries in different ways. It is therefore important for the United States to consider not only the probability of retaliatory measures, but also the effect of these measures on the relative distribution of international reserves. U.S. balance of payments measures will be more successful if they are directed against countries whose reserves have a low dollar content. From this point of view, the discrimination against Western Europe in the Interest Equalization Tax and the restrictions on capital flows were probably beneficial in reducing the drain on the U.S. gold stock.

IV. *Conclusions*

1. There is strong evidence that the reserve policies of central banks are determined to a large extent by economic factors, as the highly significant regression results show.

—The U.S. short-term investment position (official and private) is a very important determinant of reserve policies, while the return on dollar holdings plays only a minor role.

—Differences in reserve policies can largely be explained by a country's position in the world economy and, in particular, by its dependence on U.S. capital and trade. Only French reserve policies seem to have been influenced to a considerable degree by noneconomic factors.

2. The regression results highlight the urgency for the United States to eliminate her balance of payments deficit, since central banks have become reluctant to add to their nonconditional holdings of dollars.

—Past deficits will continue to cause gold outflows as a result of the dynamic nature of the adjustment process to the

new equilibrium position in central banks' portfolios.

—Future U.S. deficits will have to be paid in gold rather than be financed by increased liabilities to official institutions. Our equations indicate that further deficits may even lead to a decline in foreign exchange holdings of central banks. Thus, if the United States continues her deficits as in the past, it is not unlikely that the dollar will be devalued, although devaluation may be deferred by arranging international credits.

3. U.S. balance of payments measures have a considerable effect on the reserve policies in other countries. Possible changes in the allocation of reserves have, in turn, a significant effect on the U.S. gold stock and must therefore be taken into account when balance of payments policies are formulated.

—Balance of payments measures aimed at either repatriating short-term assets, increasing U.S. foreign liabilities, or restricting imports reduce the willingness of central banks to hold dollars. Thus, the drain on the U.S. gold stock may not be stopped, even if the U.S. balance of payments improves.

—Restrictions on U.S. long-term capital movements have comparatively little effect on reserve policies of central banks and therefore may be more effective in reducing the drain on the U.S. gold stock.

—Moreover, a cut in direct investment in the developed countries may not only improve the balance of payments, but may also reduce the apprehension of other countries at financing profitable U.S. direct investment in their own country by accumulating low-yielding dollars, thereby lessening the incentive to convert dollars into gold.

—U.S. balance of payments measures will be more successful in reducing the gold drain, if they are aimed at countries that hold relatively few dollars as international reserves.

REFERENCES

1. H. A. HAGEMANN, "Direct and Indirect Effects of the Recent British Financial Crisis on the Reserve Position of the U.S.," *Economia Internazionale*, Aug. 1965, *18*, 505–14.
2. ———, *The Supply and Demand for International Reserves*, Ph.D. thesis (unpublished), Harvard University, 1966.
3. ———, "Portfolio Selection of Central Banks and the Supply of Unconditional Reserves Under the Gold-Exchange Standard," 1966, unpublished paper.
4. P. B. KENEN, "International Liquidity and the Balance of Payments of a Reserve-Currency Country," *Quart. Jour. Econ.*, Nov. 1960, *74*, 572–86.
5. ———, *Reserve-Asset Preferences of Central Banks and Stability of the Gold-Exchange Standard*, Princeton Studies in International Finance, No. 10, Princeton, 1963.
6. INTERNATIONAL MONETARY FUND, *International Financial Statistics*.
7. —————, *Direction of Trade*.
8. U.S. DEPARTMENT OF COMMERCE, OFFICE OF BUSINESS ECONOMICS, *Survey of Current Business*, International Investment Position of the U.S.: September issues. Consistent data for earlier years were obtained directly from the Office of Business Economics, yearly data interpolated.
9. ———, *Foreign Grants and Credits by the United States Government*, reprinted in the *Statistical Abstract of the United States, 1965*.
10. U.S. TREASURY DEPARTMENT, *Treasury Bulletin*, Capital Movements, Section II, Tables 4, 5, 6, 7, 8.

More On An Empirical Definition of Money

By GEORGE G. KAUFMAN*

There is considerable disagreement among economists on the appropriate definition of money. Definitions based on apparently reasonable but empirically unsubstantiated hypotheses concerning the use of financial assets as media of exchange are neither precise nor universally accepted. A number of attempts have been made in recent years to define money empirically. Of these, three techniques appear most promising: (1) Definition by cross-elasticity of substitution of financial assets [1]; (2) Definition by discriminant analysis of time series characteristics [6]; and (3) Definition by ability to explain statistically aggregate nominal income [2]. Without evaluating the relative merits of these techniques, this paper presents additional evidence on the definition of money that best explains aggregate nominal income.

In their study for the Commission on Money and Credit, Milton Friedman and David Meiselman (F-M) used two criteria in selecting sets of financial assets to be included in the money supply: (1) the highest correlation of the sum of these assets with income and (2) higher correlation of the sum with income than of any of the components separately [2, p. 181]. The second criterion is included to ensure that an increase in correlation is attributed to the inclusion of a component in the money supply concept and not to the association between income and the particular component alone.[1]

F-M apply these dual criteria to three alternative definitions of money. They find that the set of assets, including currency, demand deposits, and time deposits at commercial banks only, satisfies the criteria better than either a narrower concept, including only currency and demand deposits, or a broader concept, including also savings-type accounts at mutual savings banks, savings and loan associations, and post offices [2, pp. 242–46].[2] The correlation for the narrower concept is consistently lower for the periods examined. The correlation for the broader concept is sometimes greater, but it is attributed to a still greater correlation between income and the additional savings-type accounts alone.

The tests described here pertain to a larger number of money supply components than considered by F-M, including a separation of currency and demand deposits. The tests also apply the F-M criteria to alternative money supply concepts observed before and after corresponding income observations as well as concurrently with income.[3] The latter

* Senior economist, Federal Reserve Bank of Chicago. The author is indebted to Anne Marie Laporte for assistance in the computations and to an unknown referee for compelling him to extract more intensively the information from the evidence.

[1] This procedure implicitly assumes some preconceived ordering of the components and does not represent a "pure" empirical test. See Robert H. Strotz [7, pp. 306–10]. In a subsequent manuscript, Milton Friedman details his ordering procedure more clearly [3]. I am indebted to Robert Laurent for initially calling my attention to this point.

[2] F-M assign each component equal weight in summing. In contrast, John G. Gurley has proposed that the components be weighed by their degree of "moneyness" [4].

[3] These tests also differ from those conducted by R. Timberlake, Jr. and J. Forston in another recent reexamination of the F-M analysis [8]. By considering the same components of the money supply as F-M but over differing observation periods, Timberlake and Forston

tests are conducted because a considerable body of doctrine suggests that money is most closely related to income in periods other than the concurrent period.[4]

All the data are seasonally adjusted quarterly first differences of logarithmic magnitudes. Income data are observations of nominal gross national product for the 1953–66 post-accord period. Money supply components are those included in the Federal Reserve Board's series on liquid assets and are observed for a slightly longer time period to maintain a fixed number of income observations in the lead-lag analysis.[5] The period is also divided into two equal subperiods (1953–59 and 1960–66) to permit examination of the stability of the estimates.

The simple correlation coefficients between income and each money component are shown in Table 1 and plotted in Figure 1 for observations of money from four quarters before the accompanying income observation through two quarters after. Marked differences among the correlation coefficients and among the correlation time patterns are readily discernible. Currency is correlated best with income in concurrent and earlier periods, and the correlation declines sharply as currency is observed before income. On the other hand, the correlation between demand deposits and income is highest when deposits precede income by a quarter and the correlation declines slowly on either side of the peak. Time deposits at commercial banks also tend to be most closely correlated with income in later periods. The highest correlations are obtained when time deposits precede income by two to four quarters. The correlation turns negative when time deposits are related to concurrent or earlier income observations.[6]

Deposits at mutual savings banks show approximately the same pattern as commercial bank time deposits, although the correlations are both lower and less stable. The correlation between savings and loan share capital and income tends to be small for almost all observations and frequently is negative. Similar to bank deposits, holdings of U.S. savings bonds are most highly correlated with income in later periods and only weakly correlated with income in earlier periods. The relationship between private nonbank holdings of marketable short-term Treasury securities and income differs from the patterns of the other financial assets and is noticeably more unstable. The pattern for the first subperiod shows these securities to be negatively correlated with income when income observations lag and positively related when income observations are concurrent or leading. The pattern for the 1960–66 subperiod indicates positive correlations except when income is observed three or more quarters later. The last panel in Figure 1 shows the correlation coefficients when time deposits at commercial banks, deposits at mutual savings banks, and share capital at savings and loan associations are combined to form a total depository savings concept. The correlation pattern of this series closely re-

conclude that little is gained by including time deposits at commercial banks in the definition of money.

[4] Although the chief proponents of a theory relating money to income in succeeding periods, F-M test for the definition of money using only concurrent observations. The highest correlation definition is then correlated with income in subsequent periods to estimate the lags.

[5] Data for money supply components are averages of figures for the last Wednesday of the three months in the quarter and the preceding month. These data for currency, demand deposits, and time deposits differ slightly from the corresponding daily average data which are unavailable for the other money supply components considered.

[6] Although not shown, the coefficients also decline when time deposits at commercial banks are correlated with income five or more quarters later. Interestingly enough, the introduction and rapid growth of negotiable CDs in the second subperiod do not appear to have altered the correlation pattern greatly although, unlike other forms of time deposits, few of these deposits are owned by individuals.

TABLE 1—SIMPLE CORRELATION COEFFICIENTS BETWEEN Δ GNP AND Δ MONEY SUPPLY COMPONENTS IN PRECEDING, SYNCHRONOUS, AND LATER PERIODS

	Quarters Money Component Leads GNP						
	+4	+3	+2	+1	0	−1	−2
				A. 1953–66			
C	−.01	−.05	.05	.20	.39	.41	.42
DD	−.01	.24	.40	.49	.44	.19	−.01
TD_B	.36	.35	.32	.20	−.09	−.24	−.17
MSB	.14	.25	.21	.17	.10	−.08	−.09
SC	−.19	−.19	−.18	−.17	−.21	−.24	−.31
SB	.23	.32	.39	.38	.25	.13	.04
US	−.31	−.46	−.39	−.23	.15	.42	.50
SAV	.34	.35	.30	.16	−.14	−.32	−.28
				B. 1953–59			
C	−.43	−.58	−.36	−.06	.36	.55	.46
DD	−.24	.03	.33	.68	.64	.24	−.14
TD_B	.19	.17	.17	.03	−.38	−.58	−.49
MSB	−.14	−.00	.19	.10	.06	−.22	−.26
SC	−.05	−.07	.03	.07	.06	−.02	−.10
SB	−.08	.07	.24	.30	.22	.01	−.16
US	−.22	−.62	−.63	−.41	.08	.51	.67
SAV	.18	.19	.24	.10	−.32	−.58	−.52
				C. 1960–66			
C	.35	.35	.30	.34	.42	.29	.41
DD	.32	.57	.58	.37	.20	.07	.06
TD_B	.53	.52	.43	.27	.02	−.19	−.16
MSB	.53	.64	.36	.33	.15	.01	−.00
SC	−.16	−.13	−.22	−.22	−.35	−.34	−.43
SB	.61	.62	.55	.41	.09	−.01	−.07
US	−.55	−.17	.21	.48	.58	.36	.13
SAV	.53	.53	.34	.17	−.10	−.27	−.30

KEY TO FIGURES AND TABLES

C = Currency outside banks
DD = Demand deposits at commercial banks
TD_B = Time deposits at commercial banks
MSB = Deposits at mutual savings banks
SC = Share capital at savings and loan associations
SB = U.S. government savings bonds plus deposits in postal savings system
US = Private nonbank holdings of U.S. government marketable securities maturing within one year
GNP = Gross national product in current dollars
$SAV = (TD_B + MSB + SC)$
$M_1 = (DD + C)$
$M_2 = (M_1 + TD_B)$
$M_3 = (M_2 + MSB + SC)$
$M_4 = (M_3 + SB)$
$M_5 = (M_4 + US)$
n = Number of quarters money leads GNP (negative values denote lags)

Note: The minimum significant r at the .05 level for the entire 1953–66 period is .25; for the subperiods the minimum significant r is .33.

sembles that of commercial bank time deposits alone, being highest when income is observed in later periods and inversely related to income in concurrent and earlier periods.

The individual components are progres-

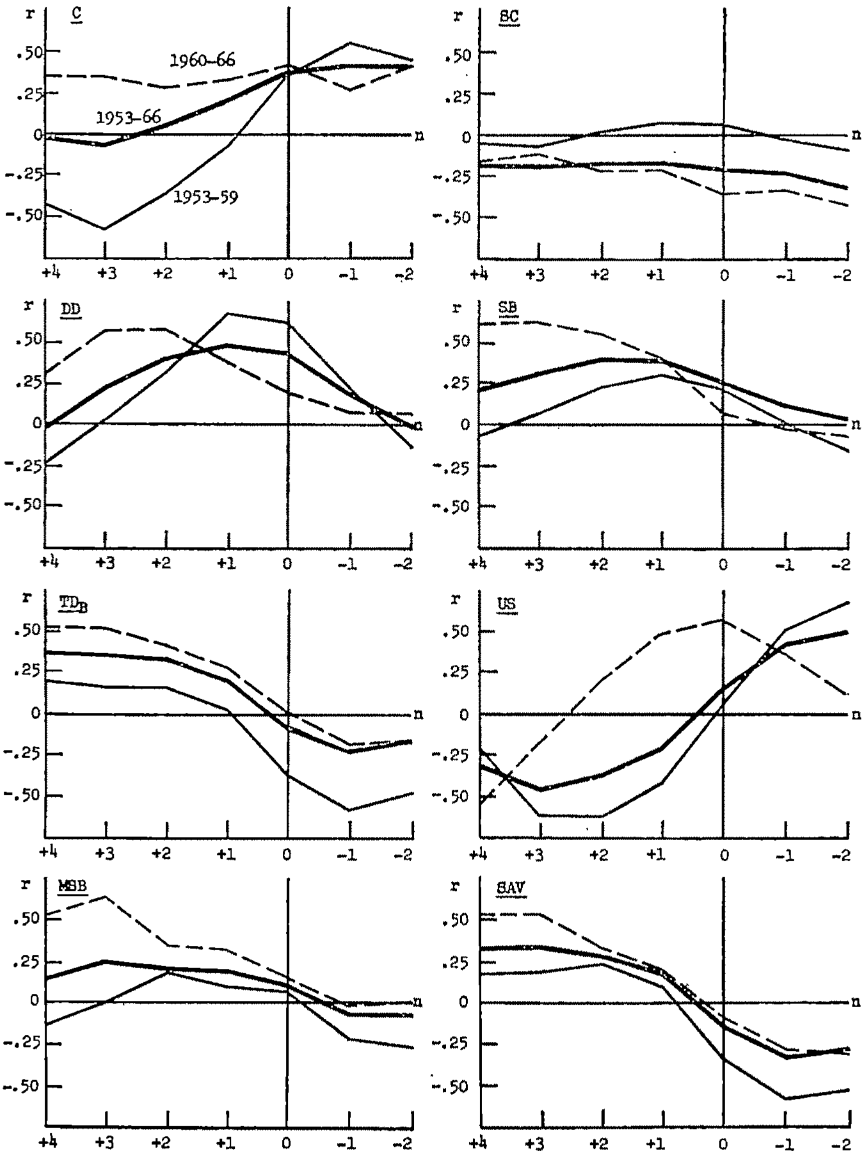


FIGURE 1—SIMPLE CORRELATION COEFFICIENTS BETWEEN Δ GNP AND Δ MONEY SUPPLY COMPONENTS

sively aggregated according to the ordering in Table 1 to form the more common alternative definitions of money. These definitions are correlated with income in the same, earlier, and later periods. The resulting correlation coefficients are shown in Table 2 and plotted in Figure 2. The correlation patterns are dominated by both demand and time deposits at commercial banks. For example, the pattern for the narrow definition of money—currency and demand deposits—resembles that for demand deposits alone, although the average time span by which the concept leads income is slightly briefer. The addition of time deposits both increases the correlation coefficients and lengthens the peak correlation to between one and

TABLE 2—SIMPLE CORRELATION COEFFICIENTS BETWEEN Δ GNP AND Δ ALTERNATIVE DEFINITIONS OF MONEY IN PRECEDING, SYNCHRONOUS, AND LATER PERIODS

	Quarters Money Definition Leads GNP						
	+4	+3	+2	+1	0	−1	−2
			A. 1953–66				
C	−.01	−.05	.05	.20	.39	.41	.42
$M_1=DD+C$	−.01	.20	.37	.49	.49*	.28	.09
$M_2=M_1+TD_B$	.25	.36*	.44*	.45	.30	.09	.03
M_2+MSB	.25	.37*	.44	.45	.30	.08	.02
$M_3=M_2+MSB+SC$	.26	.38*	.44	.44	.27	.04	−.04
$M_4=M_3+SB$	.28	.40*	.46*	.45	.29	.09	−.00
$M_5=M_4+US$	.00	.01	.12	.25	.37	.38	.37
			B. 1953–59				
C	−.43	−.58	−.36	−.06	.36	.55	.46
$M_1=DD+C$	−.33	−.11	.24	.65	.69*	.32	−.07
$M_2=M_1+TD_B$	−.13	.06	.35*	.57	.29	−.13	−.34
M_2+MSB	−.14	.06	.36*	.56	.28	−.15	−.35
$M_3=M_2+MSB+SC$	−.12	.09	.41*	.61	.33	−.13	−.35
$M_4=M_3+SB$	−.11	.12	.45*	.66	.37	−.10	−.34
$M_5=M_4+US$	−.34	−.63	−.45	−.09	.32	.47	.45
			C. 1960–66				
C	.35	.35	.30	.34	.42	.29	.41
$M_1=DD+C$	.36*	.57	.57	.40*	.28	.14	.15
$M_2=M_1+TD_B$	.53	.63*	.58	.42*	.24	.05	.11
M_2+MSB	.54*	.64	.57	.42	.24	.05	.10
$M_3=M_2+MSB+SC$	.54	.65*	.54	.35	.12	−.08	−.09
$M_4=M_3+SB$	.56	.65	.55	.38	.13	−.05	−.07
$M_5=M_4+US$	.13	.48	.65*	.58*	.39	.15	.03

* Coefficient exceeds those of any of the respective definition's components separately.

two quarters preceding income. The subsequent addition of other savings-type deposits or savings bonds neither greatly increases the correlation of the resulting money supply concepts nor greatly changes the correlation pattern. The inclusion of marketable short-term Treasury securities yields a money supply concept clearly inferior to most narrower concepts.

This evidence suggests that satisfaction of the best-fit criterion—a necessary but not sufficient condition for the satisfaction of the dual F-M criteria—depends not only on the money supply components considered but also on the range of time periods between corresponding observations of money and income. F-M's definition of money ($C+DD+TD_B$) meets this criterion better than the narrower definition only for observations two or more quarters before income, although it is not necessarily superior to broader concepts even in these periods. The narrow definition for money explains income better when observed one quarter before income through two quarters after income, although for lagging observations, currency alone is clearly superior.

The effects on the correlation coefficient of including additional money supply components to form consecutively broader concepts of money are also clearly discernible from Figures 3 and 4. Figure 3 shows the correlation pattern for each lead-lag relationship as components are progressively added to currency. The failure of components beyond time deposits at commercial banks to increase the correlation coefficients greatly becomes more evident. The unique relationship of currency is also more evident. Currency contributes little explanatory power when money leads in-

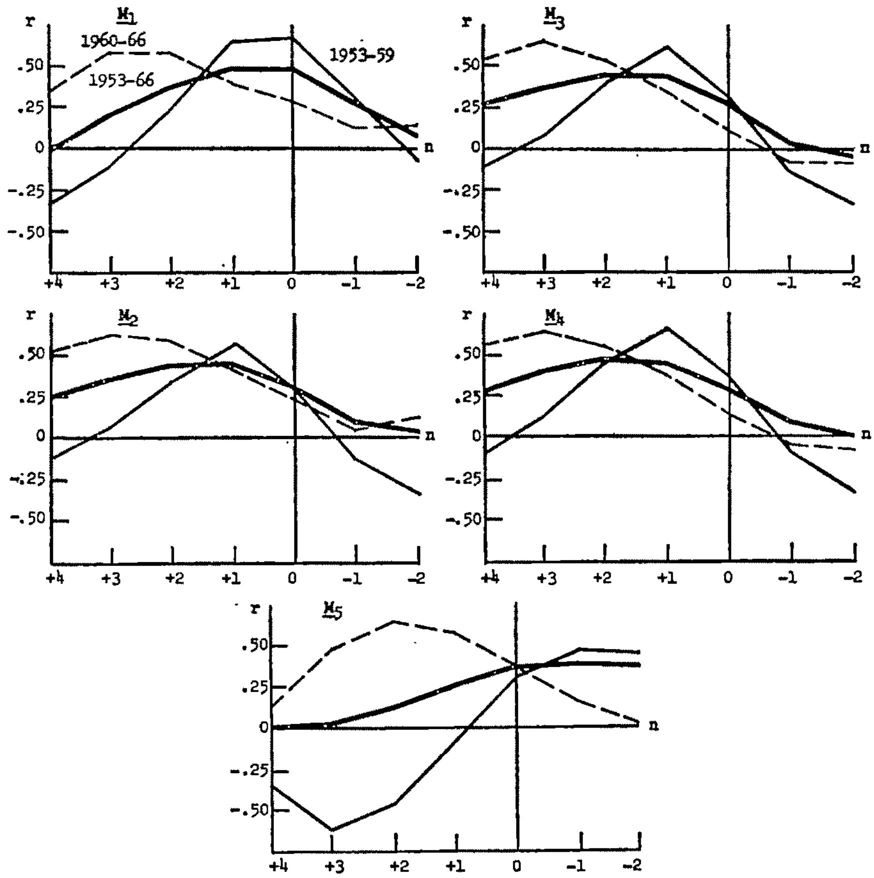


FIGURE 2—SIMPLE CORRELATION COEFFICIENTS BETWEEN Δ GNP AND Δ ALTERNATIVE DEFINITIONS OF MONEY

come and almost all the explanatory power when money is observed concurrently or after income. In Figure 4, currency is added last, rather than first. The correlations here are affected less by the progressive inclusion of additional components. For the entire period, the highest correlations are obtained when demand deposits alone are observed one quarter before income and when broader concepts are observed two quarters before income. These leads are uniformly only one quarter in the 1953–59 subperiod and, except for demand deposits alone, three quarters in the 1960–66 subperiod.

Introduction of the second of the two F-M criteria—that the correlation of the particular money supply concept with income is greater than the correlation of any of the components alone—changes the "best" definition of money derived from satisfaction of only the highest correlation criterion. A survey of Tables 1 and 2 reveals that only a limited number of definitions satisfy the second criterion. (Coefficients associated with definitions for which the income correlation exceeds that of any of its parts are marked with an asterisk in Table 2.) Concurrent observations of the narrow definition of money supply best satisfy the dual criteria for the period as a whole. Second-best solutions are obtained for a number of broader definitions of money when observed two or three quarters before income. Of these, the highest correlations are obtained for

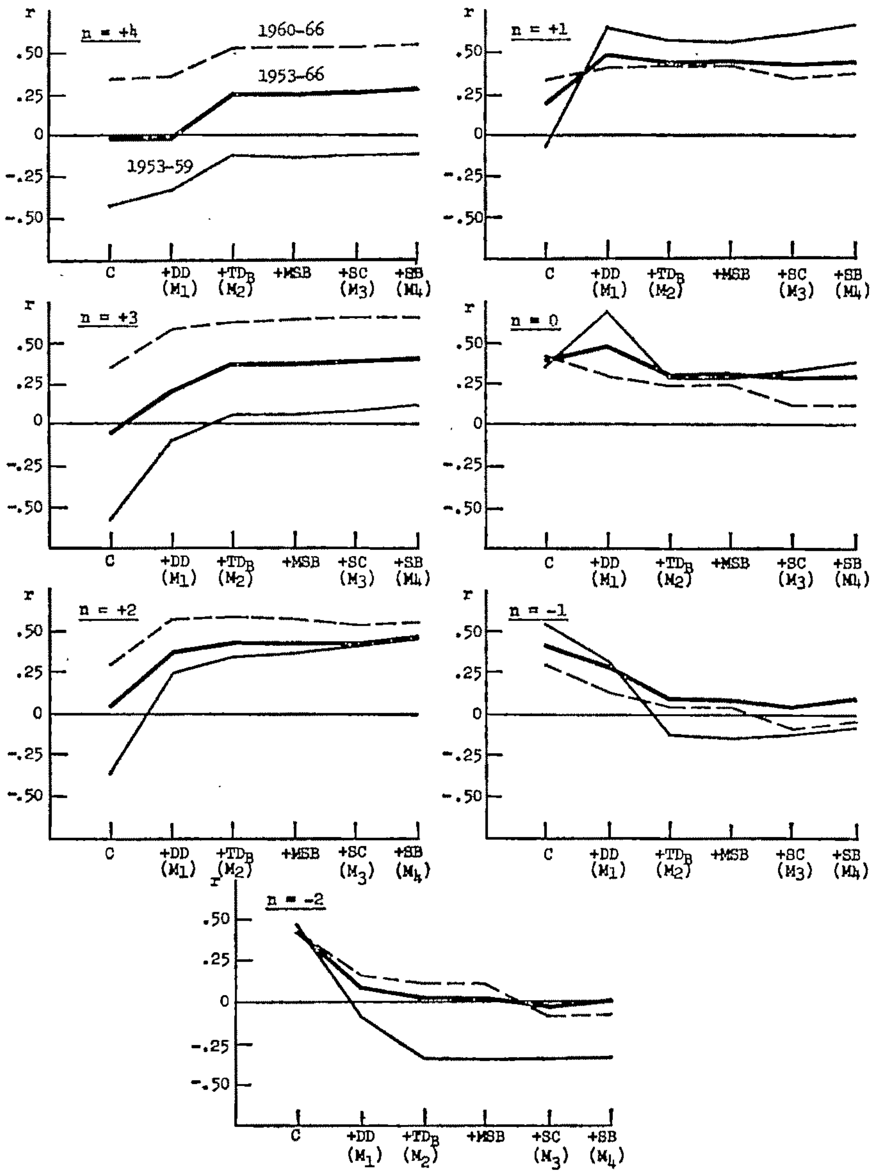


FIGURE 3—SIMPLE CORRELATION COEFFICIENTS BETWEEN Δ GNP AND Δ PROGRESSIVELY BROADER CONCEPTS OF MONEY (CURRENCY SPECIFIED FIRST)

definitions that include, in addition to currency and demand deposits, time deposits at commercial banks and those that also include all depository savings plus savings bonds.

These findings do not differ greatly for the two subperiods. Concurrent observations of the narrow definition of money satisfy the dual criteria best in the 1953–59 subperiod. However, this definition performs less well in the 1960–66 subperiod. In that period, concepts including a broader spectrum of financial assets yield the highest correlations. In addition, the time span by which the best-fit money observations precede income observations

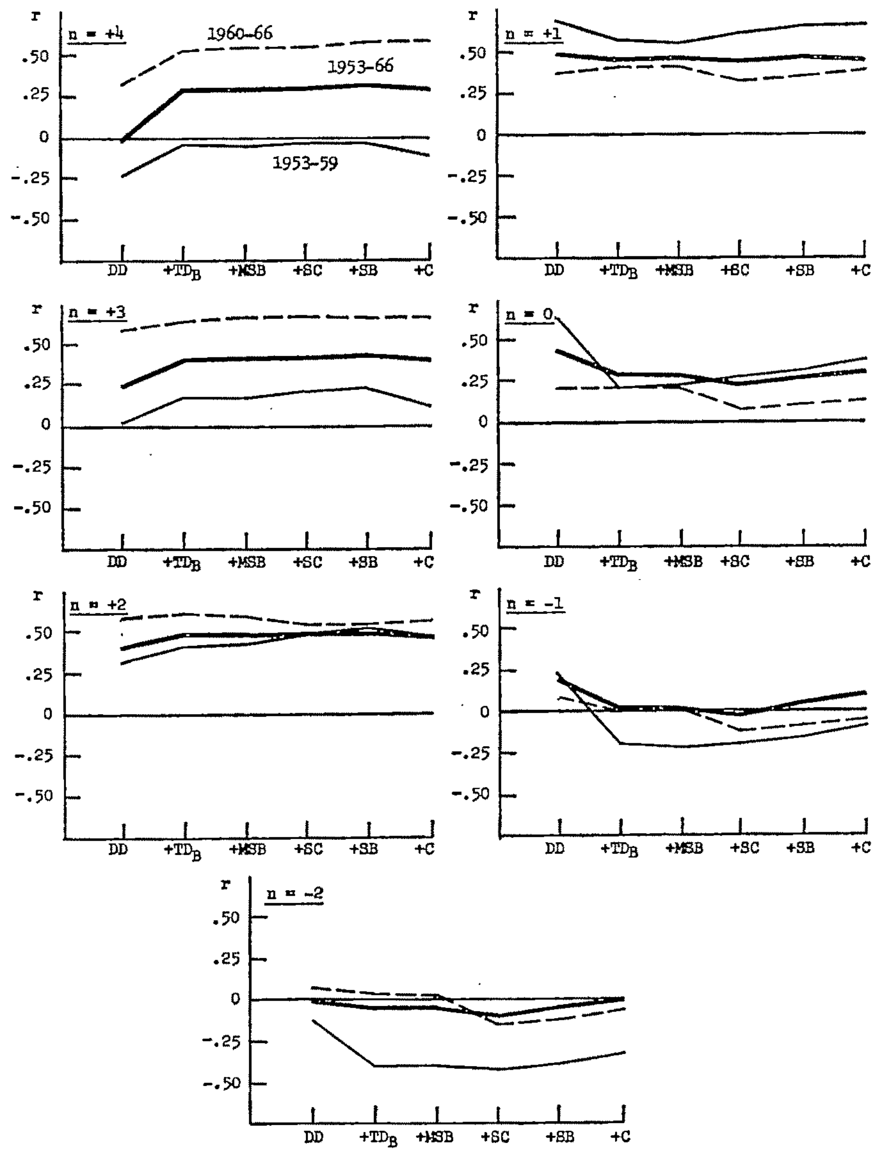


FIGURE 4—SIMPLE CORRELATION COEFFICIENTS BETWEEN Δ GNP AND Δ PROGRESSIVELY BROADER CONCEPTS OF MONEY (CURRENCY SPECIFIED LAST)

is longer in this period. Observations of the highest-correlation money concepts lead income by three quarters and observations of narrow money, by one quarter.

Although the narrow definition of money supply yields the best overall result, this definition satisfies both criteria less often than alternative concepts when the lead-lag relationships are considered individually. For the entire 1953–66 period, the narrow definition yields the best results only for concurrent observations and fails to satisfy the criteria for any other time relationships. The definition encompassing all depository savings and savings bonds in addition to currency and demand deposits performs best when money supply is observed either two or three quarters

before income. In the 1953–59 subperiod, narrow money again satisfies the criteria for concurrent observations and a broader concept for observations preceding income by two quarters. In the 1960–66 subperiod, the narrow definition fails to satisfy the criteria in any of the lead-lag observation periods, while broader concepts satisfy the criteria in each of the periods for which their observations precede those of income. Nevertheless, no single concept consistently satisfies both criteria. At times when a number of definitions satisfy the second criterion, the broadest concept generally satisfies both criteria best even though the explanatory power of the definitions is increased only slightly by the inclusion of the additional financial assets. Thus, although the F-M definition satisfies the second criterion more often than any of the other concepts, it never satisfies both. At no time do any definitions satisfy the F-M dual criteria when they are observed after the corresponding income observation.

Conclusions

The evidence developed in this study indicates that defining money according to the dual criteria established by F-M involves not only tests of alternative groupings of financial assets but also tests of alternative definitions over a number of lead-lag relationships with respect to income. Different components show different correlations as they are associated with income in preceding, concurrent, and later periods. A definition that includes demand and time deposits at commercial banks appears best at explaining income two or more quarters later. Demand deposits and currency are best at explaining income observed concurrently and one quarter later. Currency alone is the money supply concept most highly correlated with income in earlier periods. Thus, the broader the concept of money, the later are the corresponding income periods yielding the highest correlations.[7]

Because the observed correlations may be attributable to the correlation of income with a particular component of the aggregation specified rather than with the aggregation itself, F-M apply the additional criterion that the income correlation of the sum exceed the correlation of each component. Only a limited number of definitions satisfy this requirement. Overall, the narrow definition of money observed concurrently with income satisfies the dual criteria best. Broader concepts including some or all of the other financial assets satisfy the criteria more frequently when the alternative time relationships between observations of money and income are considered individually. However, no one definition satisfies the criteria consistently. Although the F-M definition including currency and all commercial bank deposits satisfies the second criterion most frequently, it never satisfies both. No definition broader than currency satisfies the criteria when observed after income.

Among other things, this evidence casts doubt on the conventional wisdom of automatically including both currency and demand deposits in any definition of money supply. Currency is more closely related to income in earlier periods, while all other components of money supply are more closely related to income in later periods.[8]

[7] These findings are not seriously altered by estimates for the earlier pre-accord 1948–52 period. Indeed, the correlation patterns are surprisingly stable, although the correlation values tend to be lower on the whole. Of the components, demand deposits are most highly correlated with income, particularly when the two series are observed concurrently. Currency again is most highly correlated with income one quarter earlier. The various savings-type funds are most highly correlated with income substantially later, but the correlations are very low.

[8] Differences in the demand functions for currency and demand deposits have been noted before. For example, see James Tobin [9], and George G. Kaufman [5]. Some observers have suggested that currency may be a closer substitute for time deposits than demand de-

The findings also indicate that although inclusion of savings-type deposits beyond time deposits at commercial banks satisfies the dual F-M criteria for the definition of money most frequently, it adds relatively little explanatory power to the definition of money. Lastly, the evidence is generally consistent with the conclusions of F-M that an important and relatively stable relationship exists between money and income in succeeding periods, although the precise characteristic of the relationship varies with the definition of money. Only when money is defined to include currency alone is support provided for a theory relating money to earlier observations of income.

REFERENCES

1. E. L. FEIGE, *The Demand for Liquid Assets: A Temporal Cross-Section Analysis*, Englewood Cliffs 1964.
2. M. FRIEDMAN AND D. MEISELMAN, "The Relative Stability of Monetary Velocity and the Investment Multiplier in the United States, 1897–1958," in Commission on Money and Credit, *Stabilization Policies*, Englewood Cliffs 1963.
3. M. FRIEDMAN AND A. J. SCHWARTZ, *Monetary Statistics of the United States*. Unpublished manuscript, National Bureau of Economic Research.
4. J. G. GURLEY, "Liquidity and Financial Institutions in the Postwar Period," Study Paper No. 14, *Study of Employment, Growth, and Price Levels*, Joint Economic Committee, 86th Cong., 1st sess., Washington 1960.
5. G. G. KAUFMAN, *The Demand for Currency*, Board of Governors of the Federal Reserve System, Staff Economic Studies, Washington 1966.
6. B. T. MCKIM, *Income Velocity and Monetary Policies in the United States: 1951–60*. Unpublished doctoral dissertation, Univ. Iowa, 1962.
7. R. H. STROTZ, "Empirical Evidence on the Impact of Monetary Variables and Aggregate Demand," in G. Horwich, ed., *Monetary Process and Policy: A Symposium*, Homewood 1967.
8. R. H. TIMBERLAKE, JR. AND J. FORSTON, "Time Deposits in the Definition of Money," *Am. Econ. Rev.*, March 1967, *57*, 190–93.
9. J. TOBIN, "Monetary Interpretation of History," *Am. Econ. Rev.*, June 1965, *55*, 464–85.

posits. The above analysis indicates, however, that the correlation patterns of currency and time deposits differ substantially more than between currency and demand deposits.

This finding has similar implications for combining currency and member bank reserves to form high powered money or the monetary base. Reserves alone tend to be most highly correlated with income in later periods. Moreover, while currency accounts for less than one-quarter of the narrow money supply and even less of broader concepts, it represents over one-half of the monetary base

Aggregate Production Functions and Types of Technical Progress: A Statistical Analysis

By MARTIN J. BECKMANN AND RYUZO SATO*

The problem of estimating the rate and type of technical progress continues to be a major concern to economists. One of the difficulties encountered is that of accurate specification both of the aggregate production function and of the type of technical progress. There are different specifications of the production function and there are different assumptions regarding the way technological advance changes the production function. In recent work the type of production function most commonly applied is the CES function. We refer specifically to the works of Arrow, *et al.* [1], David, *et al.* [3], Domar [5], Dhrymes [6], Kendrick and Sato [8], and Jorgenson and Griliches [7]. With regard to technology, exogenous or endogenous forms may be assumed. In the exogenous case which we wish to examine in this paper, it has been customary for reasons of theoretical and technical convenience to assume "neutrality" usually in the sense of Hicks or Harrod. However, as we have shown in a theoretical paper on this subject, a number of other types of "neutrality" are possible and plausible (Sato and Beckmann [10]). By "neutrality" we mean the neutrality of the effects of technical change, that is to say, relationships between certain economic variables are *invariant* under technical change [10, p. 57].

A further complication arises from the fact that these problems of specification of the form of a production function and the form of technical progress are not independent, for some forms of a production function necessarily preclude some types of technical progress. For other types of production functions certain distinctions do not arise. Consider, for example, the case of a production function of the Cobb-Douglas type. Technological change can shift the production function in any of a variety of ways, including changing the coefficients of labor and capital. But if technical change is not to change these coefficients, the change must "enter multiplicatively" and thus be either Hicks-neutral or Harrod-neutral (or their combination-factor-augmenting technical progress), but there is no way of distinguishing the two types. In other cases shown below, the choice of a specific form of a production function necessarily excludes certain types of technical progress. But the main point is that, in picking the type of technical progress, there is no reason to concentrate solely on the standard (Hicks and Harrod) forms of neutral technical progress. There are many other meaningful types of "neutral" technical progress.

In our article [10] already referred to we have systematically studied the problems mentioned above from a theoretical point of view. The paper discusses the close connection between the form of a production function in general and the type of technical progress of exogenous type, that would be consistent with a

* The authors are both professors of economics at Brown University. They wish to acknowledge the assistance of Diplomvolkswirt Lothar Weinert in performing the regressions. This work was in part supported by the National Science Foundation.

given form of a production function. It is also shown that various new types of "neutral" technical progress can be meaningfully determined and that the standard definitions of Hicks and Harrod are reclassified in a more general framework.

The present paper examines some of these results by presenting statistical evidence on the forms of production functions and on the types of technical progress. The model is applied to three countries, the United States, Japan, and Germany.

I. *The Model*

The following standard notation will be used:

Y = output
K = capital
L = labor
$y = Y/K$ = output-capital ratio
$x = L/K$ = labor-capital ratio
$z = Y/L = y/x$ = output-labor ratio
$k = K/L = 1/x$ = capital-labor ratio
r = return to capital
w = wage rate
$R = r/w$ = marginal rate of substitution
$\alpha = r/y$ = capital's relative share of income
$\beta = 1-\alpha = w/z$ = labor's relative share of income
t = time, also used as the index of exogenous technical progress.

Our general assumptions are as follows: (1) A production function of Y is homogeneous of degree one in the two-factor inputs, capital K and labor L; (2) the factor as well as commodity markets are both perfectly competitive in the long run and, hence, the factors are paid according to their marginal productivities; and (3) technical progress is purely exogenous, i.e., technical change enters in a production function simply in the form of time t.

We define a production function as:

$$Y(t) = F[K(t), L(t), t] \tag{1}$$

or alternatively,

$$y = F\left(1, \frac{L}{K}, t\right) = f(x, t) \tag{2}$$

and

$$z = F\left(\frac{K}{L}, 1, t\right) = \phi(k, t). \tag{3}$$

The marginal productivities of capital and labor are equal to their respective prices and hence

$$r = \frac{\partial F}{\partial K} = f - xf_x = \phi_k \tag{4}$$

$$w = \frac{\partial F}{\partial L} = \phi - k\phi_k = f_x. \tag{5}$$

We assume that the production function is continuous, and that the marginal productivities of both factors are positive, but that their ratio (the marginal rate of substitution) is diminishing:

$$f_x > 0 \qquad \phi_k > 0 \tag{6}$$

$$f_{xx} < 0 \qquad \phi_{kk} < 0. \tag{7}$$

In the production function (equation (1)), we introduced t as an index of the state of technology. However, in this form, the role of technology is much too general to permit a thoroughgoing analysis. It is essential to specify the way in which technology enters the production function. The usual procedure has been to formulate certain hypotheses concerning the way in which technical progress has affected relationships between certain important variables that are derived from the production function. These variables include: (1) the capital-output ratio; (2) the output per man; (3) the factor proportions; (4) the marginal productivities; and (5) the marginal rate of substitution. Thus, one might postulate that technical progress has affected any one of these characteristics in a predetermined way; for instance, that it has left a certain

variable invariant. However, since these variables will depend not only on technology but also on input proportions, it is necessary to neutralize the effect of any changes in inputs. This is the reason for considering not how variables, but how relationships between variables are affected by technical change. In particular if relationships are postulated to be invariant under technical change, one obtains the famous criteria of technical neutrality. Technical progress is Hicks-neutral when the relationship between the marginal rate of substitution and the factor proportion is unchanged. By contrast technical change is called Harrod-neutral when the relationship between the capital-output ratio and the interest rate does not change, and Solow-neutral, when the relationship between output per worker and the wage rate is invariant. In the notation above Hicks-neutrality means

$$R = R(x) \tag{8}$$

whereas with general technical progress we would have

$$R = R(x, t).$$

It is easy to show that (8) can be integrated to yield

$$Y = A(t)F(K, L)$$

where Hicks-neutral technical progress is seen to be product augmenting. In the same way we can derive the production functions which express Harrod-neutrality or Solow-neutrality.

We have generalized this concept of technical neutrality by formulating and extending this principle of invariance to all other relationships between variables: a technical progress is neutral in some sense when the relationship in which a certain economic variable stands to some other variable remains unchanged through time, that is, through technical progress (Sato and Beckmann [10]).

This principle not only yields a number of interesting new types of technical change, but suggests a more effective way of estimating production functions and technical progress: by estimating this relationship (usually a differential equation) rather than the production function in integrated form which would involve the technical progress term as a constant of integration. Of course, in making such estimates, it is necessary to specify the form of the relationships involved. We have followed the usual procedure of assuming linear and log linear relationships. (Other forms are not ruled out but would require special argument.) We examine relationships between the following seven variables: output/capital ratio; output/labor ratio; labor/capital ratio; interest rate; wage rate; marginal rate of substitution; and labor's share.

Readers of our earlier paper will note that we have dropped the elasticity of factor substitution from our list, even though it yields some of the most intriguing types of production function and of technical progress. The reason is that the approximations used to measure this elasticity from available data are too crude to allow meaningful estimates. A large percentage of the empirical elasticities of factor substitution did, in fact, turn out negative. (The presence of these errors is, in part, explained by the "impossibility theorem" of Diamond and McFadden [4].) However, there is an alternative way of introducing factor-augmenting technical progress—the *pièce de résistance* among types of progress that can be defined with the aid of the elasticity of factor substitution—e.g., by letting the interest rate be a log linear function of both the capital-output ratio and of time. In fact, there are several other similar ways.

There are $7 \cdot 6/2 = 21$ relationships that are possible between pairs of variables between which one can assume invariant

relationships. Among these, time invariant relationships between the capital-output ratio, the capital-labor ratio, and the output-labor ratio do not allow any technical change at all. Similarly an invariant relationship between pairs from the interest rate, the wage rate, and the marginal rate of substitution is inconsistent with technical change [10, p. 64]. Hicks-, Harrod-, and Solow-neutrality can each be generated in three different ways. For instance in the case of Harrod-neutrality, if the relationship between the interest rate and the capital-output ratio is invariant under technical change (Table 1, col. 1, row 1), this happens to imply that the relationship between labor's share and the output-capital ratio is also invariant (Table 1, col. 1, row 4), and that the relationship between labor's share and the interest rate is invariant (Table 1, col. 4, row 4).[1]

The types of technical progress that are thus generated fall into the following four classes:

1. *Product augmenting*

1.1 Hicks neutrality

$$Y = A(t)F(K, L)$$

where Y=output, K=capital, L=labor, and $A(t)$=technical progress term (Cases 3-6, 3-7, 6-7). Here product is increased in proportion to the output that would be obtained in the absence of technical change.

1.2 Labor additive

$$Y = A(t)L + F(K, L)$$

The increase in product is here proportional to the amount of labor used (Case 3-4).

1.3 Capital additive

$$Y = A(t)K + F(K, L)$$

The increase in product is proportional to the amount of capital used (Case 3-5).

2. *Labor augmenting*

2.1 Harrod neutrality (Cases 1-4, 1-7, 4-7)

$$Y = F(K, AL)$$

2.2 Labor combining

$$Y = F(K, A(t)K + L)$$

The augmentation of labor—as measured in efficiency units—is proportional to the amount of capital used (Case 1-5).

3. *Capital augmenting*

3.1 Solow neutrality (Cases 2-5, 2-7, 5-7)

$$Y = F(AK, L)$$

3.2 Capital combining

$$Y = F(K + A(t)L, L)$$

The augmentation of capital is proportional to the amount of labor used (Case 2-4).

4. *Input decreasing*[2]

4.1 Labor decreasing

1. The inverse production function is of the form

[1] Let the interest rate be an invariant function of the output-capital ratio. In the notation above

$$y - xy' = \phi(y), \quad \text{where } \phi \text{ denotes the invariant relationship.}$$

Rearranging terms

$$\frac{xy'}{y} = \frac{y - \phi(y)}{y} = \psi(y) \quad \text{say.}$$

This says that labor's share is an invariant function of the output-capital ratio.

Eliminating y between the above two relations we obtain

$$\frac{xy'}{y} = \xi(y - xy')$$

i.e., the relationship between the share and the interest rate is invariant.

[2] Readers should note that following the suggestion of the referee we have changed Anti-Hicks in the previous article [10, p. 60] to Input Decreasing.

TABLE 1—TYPES OF TECHNICAL PROGRESS IMPLIED BY ALTERNATIVE INVARIANT RELATIONSHIPS

	1 $\frac{\text{output}}{\text{capital}}$ y	2 $\frac{\text{output}}{\text{labor}}$ z	3 $\frac{\text{labor}}{\text{capital}}$ x	4 interest r	5 wage w	6 MRS R
4 interest r	Harrod II	capital combining V	labor additive IX			
5 wage w	labor combining IV	Solow III	capital additive VIII	No		
6 MRS R	labor VI decreasing	capital VII decreasing	Hicks I	technical progress		
7 share β	Harrod II	Solow III	Hicks I	Harrod II	Solow III	Hicks I

$$L = G(K, Y) + C(t)Y \qquad \frac{dC}{dt} < 0$$

where $C(t)$ is decreasing with time. The reduction of the labor input is thus proportional to output Y (Case 1-6).

4.2 Capital decreasing

2. This is the symmetric case when the inverse production function is of the form

$$K = H(K, Y) + C(t)Y \qquad \frac{dC}{dt} < 0$$

Here the reduction of capital is proportional to output (Case 2–6).

Thus, technical progress may operate on the input side or on the product side and its effect may be proportional to any of the basic magnitudes: output, labor, or capital input. Combinations of these types may also arise. The invariance principle will generate certain of these when the elasticity of factor substitution is added to our list of variables, so that second order differential equations arise which involve two time terms as "constants" of integration.

Alternatively, such combinations arise when a time term is entered into the postulated relationships to express the null hypothesis that the relationship is *not* invariant with respect to time. The most important of these combinations is:

5. *Factor augmenting technical progress*

$$Y = F[A(t)K, B(t)L]$$

This combination of Harrod and Hicks, or Solow and Hicks neutrality is obtained in Case 1-7 when the capital-output ratio is a separable function of labor's share and of time[8]

$$\frac{Y}{K} = A(t)\phi(\text{share}).$$

Among the types of technical progress that we have obtained in this way are several which are well known in the literature: Hicks, Harrod, Solow, and factor augmenting technical change. The less obvious new types are all of the kind where the augmentation (reduction) of one variable is proportional to some other variable.

II. *Statistical Analysis*

We have applied regression analysis to time series data for the U.S., Japanese, and German private nonfarm sectors (see the table in the Appendix). Linear and log-linear regression equations are estimated for the relationships in Table 1—specifica-

[8] Factor-augmenting technical progress can be obtained in a more general way from the invariant relationship between the share and the elasticity of factor substitution (See our article [10, p. 63]).

tion in terms of two economic variables—rather than from the implied production functions. The direct estimation of production functions is in general more difficult because of nonlinearity of the functions. Also by using these relationships it becomes unnecessary to estimate the constants of integration that enter the production functions.

It should be noted that by specifying the alternative formulations as linear or log-linear functions we have at the same time specified the form of the production functions. For example, to test Harrod neutrality we applied,

$$(9) \quad r = a + by + u(t) \qquad r = \text{interest rate}, \quad y = \frac{\text{output}}{\text{capital}}$$

and

$$(10) \quad \log r = a + b \log y + u(t)$$

where $u(t)$ is a random variable satisfying $E[u]=0$ and $E[uu']=\sigma^2 I$. By assuming r as a linear function of y, we have specified that the production function must be equal to:

$$y = \frac{1}{1-b}(A(t)x)^{1-b} + \frac{a}{1-b}, \quad 0 < b < 1,$$

which is a combination of the Cobb-Douglas type with a linear function in terms of K.[4] This follows by integration of the differential equation (9). Also the assumption of a log-linear form, $\log r = a + b \log y$, implies a production function of the CES type:

$$y = [\alpha + (A(t)x)^{1-b}]^{1/(1-b)}, \quad \log \alpha = a$$

Table 2 summarizes the implied specification of production functions corresponding to a linear and (whenever integration is possible) log-linear alternative form of alternative specifications.

[4] This function is the same as the one obtained in the article [12] as a CEDD (constant elasticity of derived demand) production function.

The overall conclusion to be drawn from Table 2 is that in virtually every case the implied production function turns out to be a modification of the Cobb-Douglas or CES production functions. In fact, the estimated values of the coefficients in these functions tend to make the approximation to Cobb-Douglas usually quite close. This is yet another confirmation of the robustness of the Cobb-Douglas function for empirical work. It also removes to a large extent the difficulty or controlling the form of the production function while testing for the type of technical progress, since the production function turns out to be very much the same in every case. Certain of the new functional forms obtained may be of interest in themselves as generalizations of Cobb-Douglas and/or CES and may merit further empirical exploration.

III. *Results*

The linear and log-linear equations in the second column of Table 2 were estimated in the form:

$$q(t) = a + bp(t) + u(t) \quad \text{or}$$
$$\log q(t) = a + b \log p(t) + u(t),$$

where $p(t)$ and $q(t)$ stand for the economic variables listed in Table 2, such as R and x for the case of Hicks neutrality. To test the invariance of these relationships with respect to technical change, i.e., time, we compared the fit as measured by R^2 with that obtained when a time term is added:

$$q(t) = a + bp(t) + ct + u(t) \quad \text{or}$$
$$\log q(t) = a + bp(t) + ct + u(t).$$

In Table 3, the values of R^2 are computed and ranked for different definition of "neutrality" in three countries. It is interesting to note that Solow Neutral Technical Progress ranks first in both the United States and Germany, and third in Japan. Harrod neutrality ranked

TABLE 2—IMPLIED PRODUCTION FUNCTIONS CORRESPONDING TO LINEAR (OR LOG-LINEAR) FORMS OF ALTERNATIVE SPECIFICATIONS

Types	Forms	Alternative Specifications	Implied Production Functions
I. Hicks	Linear	$R = a + bx$	$y = A(t)[a + (1+b)x]^{1/(1+b)}$
	Log-linear	$\log R = a + b \log x$	$y = A(t)[\alpha + x^{1-b}]^{1/(1-b)}$, $\log \alpha = a$ CES $b = 1$ (United States and Germany) $y = A(t)x^{1/(1+\alpha)}$ CD $b = 3$ (Japan), $y = A(t)\dfrac{x}{\sqrt{1+\alpha x^2}}$
II. Harrod	Linear	$r = a + by$	$y = \dfrac{(A(t)x)^{1-b}}{1-b} + \dfrac{\alpha}{1-b}$, $0 < b < 1$
	Log-linear	$\log r = a + b \log y$	$y = [\alpha + (A(t)x)^{1-b}]^{1/(1-b)}$
III. Solow	Linear	$w = a + bz$	$z = \dfrac{(A(t)k)^{1-b}}{1-b} + \dfrac{\alpha}{1-b}$, $0 < b < 1$
	Log-linear	$\log w = a + b \log z$	$z = [\alpha + (A(t)k)^{1-b}]^{1/(1-b)}$
IV. Labor-combining	Linear	$w = a + by$	$y = A(t)e^{bx} - \dfrac{a}{b}$, $b < 0$,
	Log-linear	$\log w = a + b \log y$	$y = [\alpha(1-b)x + A(t)]^{1/(1-b)}$ $\log \alpha = a$, $b < 0$ (Germany and Japan) $b \neq 1$
V. Capital-combining	Linear	$r = a + bz$	$z = A(t)e^{bk} - \dfrac{a}{b}$, $b < 0$,
	Log-linear	$\log r = a + \beta \log z$	$z = [\alpha(1-b)k + A(t)]^{1/(1-b)}$, $b \neq 1$, $\log \alpha = a$
VI. Labor-decreasing	Linear	$R = a + by$	$x = -a + A(t)y + by \log y$
	Log-linear	$\log R = a + b \log y$	$x = \dfrac{\alpha}{b-1} y^{b-1} + A(t)y$, $b \neq 1$ $\log \alpha = a$
VII. Capital-decreasing	Linear	$R = a + bz$	$k = -a + A(t)z + bz \log z$
	Log-linear	$\log R = a + b \log z$	$k = \dfrac{\alpha}{b-1} z^{b-1} + A(t)z$, $\beta \neq 1$ $\log \alpha = a$
VIII. Capital-additive	Linear	$w = a + bx$	$y = ax + \dfrac{b}{2}x^2 + A(t)$
	Log-linear	$\log w = a + b \log x$	$y = \dfrac{\alpha}{1+b} x^{1+b} + A(t)$, $\log \alpha = a$
IX. Labor-additive	Linear	$r = a + bk$	$z = ak + \dfrac{b}{2}k^2 + A(t)$
	Log-linear	$\log r = a + b \log k$	$z = \dfrac{\alpha}{1+b} k^{1+b} + A(t)$, $\log \alpha = a$

first and second in Japan and the United States, but did not rank well in Germany (Rank 7). Hicks neutrality does not rank as high as the Solow or Harrod types. Capital-Additive Neutrality ranks very high in Germany (Rank 2), but not in the United States and in Japan.

It is understandable to have such high

TABLE 3—COMPARISONS OF R^2 IN LOG-LINEAR REGRESSIONS AMONG THREE COUNTRIES

Type	United States		Japan		Germany	
	R^2	Rank	R^2	Rank	R^2	Rank
I. Hicks	.831	4	.785	2	.708	4
II. Harrod	.933	2	.855	1	.422	7
III. Solow	.944	1	.758	3	.980	1
IV. Labor-combining	.897	3	.021	8	.770	3
V. Capital-combining	.818	5	.039	7	.272	9
VI. Labor-decreasing	.466	8	.755	4	.692	5
VII. Capital-decreasing	.702	7	.001	9	.653	6
VIII. Capital-additive	.779	6	.473	6	.950	2
IX. Labor-additive	.411	9	.633	5	.347	8

ranks in the Solow and Harrod neutral cases, since the log-linear relations imply that the production functions are of the CES type. Thus, Solow and Harrod neutral types are cast in a particularly favorable position. In order to offset this effect of the form of the production function, we might consider imbedding the various types of neutrality in CES functions and covering these estimates. However, this method becomes feasible only as further progress is made in nonlinear econometric estimation.

The above observation (Table 3) is made by comparing only the log-linear estimates. Our findings may now be summarized as follows:

(1) Among the unconventional types of neutral technical progress several perform well in one or the other country, but none is leading. In the United States, labor-combining and capital-combining technology; in Japan, labor-decreasing and labor-additive technology; in Germany, both labor and capital decreasing, and capital-additive technology. Such a phenomenon, however, might well be expected when fitting a number of different forms to different time series.

(2) The traditional types of Hicks, Harrod, and Solow neutrality are for all countries at least as good as the unconventional types of neutrality. It is noteworthy that Solow (i.e., capital augmenting) technical change performs particularly well.

(3) When in the equations for Hicks, Harrod, and Solow neutrality, time is entered as an additional explanatory variable, a closer fit can be expected. Interestingly enough, this effect is quite small. We conclude that general factor-augmenting technical progress does not give a substantially improved explanation of observed data when compared with technical progress that augments only one factor.

(4) No matter how technical change is specified, the estimated production function turns out to be close to a Cobb-Douglas or CES function in every case.

TABLE 4—SUMMARY OF REGRESSION ANALYSIS

Case	Form	United States				Japan				Germany			
		a	b	c	R^2	a	b	c	R^2	a	b	c	R^2
I	A	− .026,389	.580,145 (.039,535)		.811,588	− .254,093	.796,724 (.046,880)		.908,754	.058,087	.276,869 (.008,406)		.926.549
	B	− .485,458	1.198,325 (.076,493)		.830,747	− .383,005	2.847,489 (.277,063)		.784,587	−1.380,571	1.085,609 (.075,132)		.708,263
	C	.013,551	.504,116 (.090,435)	−.000,322 (.000,344)*	.814,856	− .375,492	.858,207 (.095,078)	.001,805 (.002,420)*	.910,532	.817,057	.174,467 (.028,842)	−.009,775 (.002,649)	.936,691
	D	− .538,890	1.107,802 (.172,822)	−.001,026 (.001,759)*	.831,920	−2.331,131	4.005,844 (.433,596)	.055,876 (.017,387)	.842,632	3.475,582	−1.008,516 (.541,317)	−.069,990 (.017,942)	.752,562
II	A	.013,864	.302,251 (.012,080)		.926,041	− .030,018	.479,584 (.039,861)		.833,096	.001,823	.224,966 (.014,461)		.737,879
	B	−1.168,961	.911,675 (.034,449)		.933,367	1.793,163	2.693,458 (.206,203)		.854,724	−1.518,591	1.295,658 (.163,260)		.422,751
	C	.009,807	.323,825 (.032,859)	−.000,149 (.000,211)	.926,784	.017,492	.396,445 (.035,632)	−.000,789 (.000,172)	.904,616	.052,942	.189,708 (.042,873)	−.000,442 (.000,506)*	.740,153
	D	−1.106,066	.955,426 (.095,211)	−.000,710 (.001,438)*	.933,696	1.966,091	2.617,725 (.228,667)	−.007,194 (.009,116)*	.857,885	−1.466,777	1.181,552 (.496,687)	−.007,604 (.006,590)*	.423,153
III	A	− .036,839	.697,252 (.007,673)		.993,982	− .026,839	.875,767 (.075,674)		.822,242	.010,752	.746,239 (.014,016)		.970,554
	B	− .414,634	1.031,527 (.011,106)		.944,237	− .137,289	1.103,453 (.115,775)		.758,013	− .265,134	.994,800 (.015,308)		.980,042
	C	− .063,478	.821,924 (.033,146)	−.003,455 (.000,908)	.995,374	− .069,000	.577,292 (.077,555)	.002,451 (.000,457)	.912,358	.002,807	.654,477 (.027,732)	.001,058 (.000,281)	.974,740
	D	− .297,913	1.196,377 (.086,169)	−.003,878 (.002,011)*	.994,642	−1.407,621	.705,785 (.109,861)	.014,953 (.002,745)	.882,517	− .642,899	.801,519 (.066,744)	.004,132 (.001,392)	.981,917
IV	A	− .397,308	2.771,952 (.155,091)		.864,662	.208,065	− .242,226 (.213,943)		.042,331	.763,165	− .487,095 (.046,784)		.557,616
	B	1.021,123	1.514,068 (.072,189)		.897,937	−2.025,586	− .115,743 (.145,940)		.021,229	−1.500,712	−1.385,858 (.081,748)		.769,682
	C	− .000,182	.660,232 (.272,677)	.014,598 (.001,752)	.944,028	.141,036	.368,657 (.112,215)	.005,789 (.000,542)	.811,569	− .771,899	.571,675 (.067,269)	.013,280 (.000,794)	.896,873
	D	− .657,637	.346,295 (.088,964)	.018,950 (.001,344)	.979,812	−2.747,259	.200,311 (.080,733)	.030,022 (.003,218)	.761,732	−2.413,575	.624,482 (.092,287)	.028,264 (.001,224)	.968,307
V	A	.067,161	.064,158 (.005,125)		.755,093	.054,108	− .065,174 (.079,121)		.022,862	.281,014	.184,381 (.026,082)		.367,537
	B	−2.026,375	.552,497 (.036,890)		.817,724	−2.189,412	.919,042 (.846,102)		.039,094	−2.404,576	− .661,721 (.116,591)		.272,492
	C	.078,873	.009,346 (.023,955)	.001,533 (.000,656)	.782,372	.106,872	.308,362 (.060,601)	−.003,068 (.000,357)	.731,342	.311,397	.166,519 (.035,548)	−.004,044 (.000,358)	.746,764
	D	1.914,342	.710,726 (.295,951)	−.003,722 (.006,907)	.818,795	7.889,027	4.074,027 (.722,973)	−.118,631 (.018,064)	.621,728	3.159,768	2.185,240 (.430,130)	−.060,864 (.008,971)	.528,056

TABLE 4 (continued)

Case	Form	United States				Japan				Germany			
		a	b	c	R^2	a	b	c	R^2	a	b	c	R^2
VI	A	.305,742	− .262,962 (.039,162)		.474,163	− .244,581	3.539,219 (.487,176)		.645,376	.966,664	2.131,526 (.073,539)		.907,141
	B	−2.190,085	− .652,395 (.091,272)		.465,582	3.818,745	2.809,199 (.297,411)		.754,690	− .017,879	2.681,516 (.192,718)		.692,422
	C	.210,313	.244,507 (.073,321)	−.003,508 (.000,471)	.753,400	.442,668	2.336,622 (.364,200)	−.011,413 (.001,760)	.858,254	.564,030	1.075,771 (.182,100)	−.013,242 (.002,150)	.935,798
	D	− .448,352	.609,184 (.171,508)	−.019,661 (.002,592)	.754,245	4.713,351	2.417,409 (.282,963)	−.037,217 (.011,280)	.823,361	.946,798	.557,070 (.533,273)	−.029,864 (.007,075)	.745,722
VII	A	.282,890	− .075,764 (.006,762)		.715,145	.577,124	−1.360,882 (.621,724)		.141,789	1.774,939	−1.972,929 (.182,406)		.576,330
	B	−1.611,741	− .479,029 (.044,105)		.702,315	−2.052,128	− .184,414 (.957,427)*		.001,278	−2.139,442	−1.656,522 (.130,218)		.652,985
	C	.279,586	− .060,345 (.033,245)*	−.000,432 (.000,910)*	.716,453	.981,497	1.501,815 (.505,887)	−.023,509 (.002,980)	.733,693	2.029,136	.962,844 (.146,791)	−.033,832 (.001,492)	.939,873
	D	−1.616,149	− .485,254 (.354,881)*	.000,146 (.008,282)*	.702,315	9.296,663	3.368,245 (.824,213)	−.133,584 (.020,594)	.600,941	3.802,668	1.383,721 (.491,430)	−.064,996 (.010,249)	.764,430
VIII	A	2.426,093	−4.328,043 (.361,632)		.741,247	.272,995	− .152,912 (.030,229)		.468,743	.532,799	− .064,474 (.005,779)		.591,430
	B	−2.368,070	−2.099,567 (.158,346)		.778,575	−2.060,528	− .542,902 (.106,499)		.472,600	.750,778	− .615,976 (.015,408)		.948,940
	C	.020,847	.250,605 (.409,138)*	.019,397 (.001,558)	.937,812	− .138,094	.055,285 (.042,097)	.006,113 (.001,072)	.754,092	− .755,220	.109,309 (.008,507)	.016,589 (.000,781)	.935,175
	D	−1.121,793	.011,833 (.124,014)*	.023,930 (.001,258)	.973,577	−3.028,126	.032,434 (.144,630)*	.027,753 (.005,800)	.709,868	−1.565,577	− .264,609 (.114,248)*	.011,743 (.003,787)	.954,130
IX	A	.024,233	.043,923 (.008,179)		.365,768	.100,242	− .035,127 (.002,382)		.882,303	.265,131	− .102,244 (.012,358)		.443,169
	B	−2.853,472	.901,195 (.152,493)		.411,244	−2.443,550	−2.304,631 (.325,759)		.633,148	−2.131,350	− .469,633 (.069,528)		.346,625
	C	.189,751	− .058,582 (.006,107)	.003,222 (.000,170)	.924,151	.094,765	− .037,617 (.003,797)	.000,218 (.000,258)*	.885,229	.329,986	.070,766 (.020,959)	−.003,841 (.000,422)	.718,105
	D	−1.660,905	−1.119,328 (.124,636)	.022,902 (.001,265)	.923,467	−5.359,488	−4.038,470 (.447,677)	.083,634 (.017,951)	.793,343	1.910,007	1.273,127 (.509,318)	−.058,247 (.016,886)	.426,858

Note: 1. The values in parentheses are t statistics.
2. ( )* indicates that the coefficient is not significantly different from zero.
3. Form A $\quad q = a+bp+u$
B $\log q = a+b \log p+u$
C $\quad q = a+bp+cl+u$
D $\log q = a+b \log p+cl+u$
4. The data in the appendix are constructed by the authors from the articles [11] [12] and also from W. G. Hoffmann, *Das Wachstum der deutschen Wirtschaft seit der Mitte des 19. Jahrhunderts*, New York 1965.

APPENDIX

Time Series for the United States, Japan, and German Federal Republic Private Non-Farm Sectors

	United States (Y and K are measured at 1929 dollars)					Japan (Y and K are measured at 1930 Yen)				
Year	$y=\frac{Y}{K}$	$z=\frac{Y}{L}$	$x=\frac{L}{K}$	r	w	$y=\frac{Y}{K}$	$z=\frac{Y}{L}$	$x=\frac{L}{K}$	r	w
1909	.3226	.6710	.4808	.108	.446					
1910	.3086	.6525	.4730	.102	.437					
1911	.3229	.6732	.4796	.108	.448					
1912	.3140	.6626	.4739	.104	.444					
1913	.3315	.7034	.4713	.112	.468					
1914	.3120	.6444	.4842	.101	.435					
1915	.3079	.6587	.4674	.106	.432					
1916	.3286	.7216	.4553	.118	.463					
1917	.2999	.6632	.4523	.111	.418					
1918	.3131	.7200	.4349	.107	.474					
1919	.3267	.7821	.4177	.116	.505					
1920	.3241	.7825	.4142	.103	.533					
1921	.3765	.8611	.4373	.139	.543					
1922	.3680	.8359	.4403	.125	.553					
1923	.3616	.8694	.4159	.122	.576					
1924	.3760	.9380	.4009	.124	.628					
1925	.3573	.9157	.3902	.120	.608					
1926	.3584	.9447	.3794	.117	.636					
1927	.3592	.9460	.3797	.116	.640					
1928	.3531	.9542	.3700	.107	.632					
1929	.3551	.9899	.3588	.118	.661					
1930	.3288	.9751	.3372	.114	.637	.1846	.200	.9229	.0593	.1358
1931	.3387	1.0088	.3357	.110	.681	.1896	.206	.9202	.0531	.1483
1932	.3307	.9774	.3383	.131	.589	.1871	.198	.9451	.0575	.1372
1933	.3314	.9432	.3514	.120	.602	.1891	.195	.9697	.0594	.1338
1934	.3581	1.0429	.3434	.127	.673	.2030	.206	.9854	.0654	.1397
1935	.4018	1.1242	.3574	.141	.730	.1942	.201	.9663	.0689	.1296
1936	.4254	1.1342	.3751	.152	.729	.1925	.203	.9482	.0670	.1324
1937	.4374	1.1717	.3733	.149	.773	.2000	.216	.9259	.0800	.1296
1938	.4346	1.1885	.3656	.144	.795	.1874	.210	.8922	.0665	.1355
1939	.4594	1.2183	.3771	.159	.796	.1787	.204	.8758	.0609	.1344
1940	.4812	1.2713	.3785	.172	.817	.1612	.196	.8225	.0584	.1250
1941	.5013	1.3040	.3845	.189	.812	.1665	.207	.8044	.0574	.1356
1942	.5007	1.3130	.3813	.178	.846	.1490	.196	.7605	.0525	.1270
1943	.5163	1.3408	.3850	.177	.882	.1390	.192	.7241	.0489	.1244
1944	.5554	1.4629	.3796	.184	.977	.1180	.179	.6590	.0408	.1171
1945	.5667	1.5282	.3708	.178	1.048	.1025	.192	.5337	.0349	.1265
1946	.5497	1.4512	.3788	.172	.998	.0540	.123	.4392	.0032	.1156
1947	.4320	1.4192	.3749	.174	.955	.0595	.160	.3720	.0011	.1570
1948	.5372	1.4797	.3630	.178	.988	.0698	.195	.3580	.0021	.1892
1949	.5228	1.5201	.3439	.170	1.025	.0815	.230	.3542	.0101	.2015
1950	.5537	1.6269	.3403	.201	1.036	.0984	.266	.3700	.0160	.2226
1951	.5459	1.6545	.3299	.188	1.084	.1066	.247	.4316	.0193	.2023
1952	.5380	1.6847	.3193	.171	1.151	.1148	.261	.4397	.0165	.2234
1953	.5427	1.7058	.3182	.169	1.175	.1191	.267	.4461	.0197	.2229
1954	.5357	1.7820	.3006	.163	1.238	.1193	.256	.4659	.0175	.2184
1955	.5542	1.8618	.2977	.182	1.249	.1275	.271	.4705	.0199	.2287
1956	.5457	1.6546	.3298	.174	1.127	.1414	.293	.4827	.0266	.2379
1957	.5394	1.8912	.2852	.168	1.303	.1452	.291	.4989	.0184	.2540
1958	.5358	1.9360	.2768	.161	1.353	.1482	.294	.5040	.0280	.2384
1959	.5506	1.9993	.2754	.174	1.368	.1659	.342	.4850	.0390	.2616
1960	.5520	2.0342	.2714	.171	1.406	.1825	.380	.4802	.0478	.2804

GERMAN FEDERAL REPUBLIC

Year	Y (GNP) Prices of 1913 (Million Mark)	K Capital Stock Prices of 1913 (Mrd. Mark)	L Labor Force in Hours Worked (Billion hours)	$\beta=1-\alpha$ Share of Labor Income
1850	9,449	7.16	56.7	.819
1851	9,390	7.19	56.7	.819
1852	9,578	7.38	56.7	.819
1853	9,565	7.50	56.7	.819
1854	9,793	7.63	56.7	.778
1855	9,657	7.79	58.4	.778
1856	10,442	7.98	58.4	.778
1857	10,948	8.10	58.4	.778
1858	10,888	8.33	58.4	.778
1859	10,938	8.48	58.4	.778
1860	11,577	8.65	58.4	.753
1861	11,364	8.89	58.4	.753
1862	11,872	9.22	58.4	.753
1863	12,729	9.55	58.4	.753
1964	13,127	8.85	58.4	.753
1865	13,167	10.00	58.4	.750
1866	13,293	10.20	58.4	.750
1867	13,318	10.30	60.9	.750
1868	14,099	10.60	60.9	.750
1869	14,188	10.90	60.9	.750
1870	14,169	11.70	60.9	.778
1871	14,653	12.50	60.9	.778
1872	15,683	13.10	60.9	.778
1873	16,347	13.70	60.9	.778
1874	17,545	14.20	60.9	.778
1875	17,651	14.60	68.8	.800
1876	17,548	14.80	68.8	.800
1877	17,438	15.00	68.8	.800
1878	18,257	15.20	68.8	.800
1879	17,839	15.50	68.8	.800
1880	17,679	16.05	68.8	.766
1881	18,122	16.80	68.8	.766
1882	18,441	17.40	68.8	.766
1883	19,427	18.50	68.8	.766
1884	19,923	19.70	68.8	.766
1885	20,417	20.80	73.7	.762
1886	20,548	22.20	73.7	.762
1887	21,362	23.50	73.7	.762
1888	22,266	25.00	73.7	.762
1889	22,859	26.60	73.7	.762
1890	23,589	28.30	73.7	.741
1891	23,579	29.20	73.7	.741
1892	24,539	29.80	73.7	.741
1893	25,760	31.30	73.7	.741
1894	26,383	32.90	73.7	.741
1895	27,621	34.60	80.7	.731
1896	28,615	37.40	80.7	.731
1897	29,437	40.50	80.7	.731
1898	30,703	44.40	80.7	.731
1899	31,813	47.50	80.7	.731

German Federal Republic (Continued)

Year	Y (GNP) Prices of 1913 (Million Mark)	K Capital Stock Prices of 1913 (Mrd. Mark)	L Labor Force in Hours Worked (Billion hours)	$\beta=1-\alpha$ Share of Labor Income
1900	33,169	49.80	80.7	.726
1901	32,406	51.20	80.7	.726
1902	33,142	52.00	80.7	.726
1903	34,979	54.00	80.7	.726
1904	36,405	57.00	80.7	.726
1905	37,189	60.10	80.7	.712
1906	38,283	63.40	90.2	.712
1907	39,993	67.00	90.2	.712
1908	40,665	69.10	90.2	.712
1909	41,482	71.80	90.2	.712
1910	42,981	74.30	90.2	.709
1911	44,478	77.90	90.2	.709
1912	46,388	82.00	90.2	.709
1913	48,480	85.20	99.0	.709
1925	45,515	76.63	89.0	.873
1926	43,688	77.63	89.0	.873
1927	51,806	81.92	89.0	.873
1928	52,969	85.40	89.0	.873
1929	53,596	86.60	89.3	.873
1930	50,326	86.70	89.3	.971
1931	45,226	83.81	89.3	.971
1932	41,011	82.27	89.3	.971
1933	45,068	82.41	89.3	.971
1934	49,395	83.96	89.3	.971
1935	53,856	86.62	89.3	.781
1936	59,511	88.86	89.3	.781
1937	63,098	96.03	89.3	.781
1938	67,967	103.22	92.9	.781
1950	40,052	69.54	52.9	.740
1951	44,151	74.03	52.9	.740
1952	46,278	78.98	52.9	.740
1953	49,090	83.28	52.9	.740
1954	53,828	89.31	52.9	.740
1955	59,393	97.54	52.9	.728
1956	62,283	104.86	52.9	.728
1957	64,363	112.25	52.9	.728
1958	66,613	118.87	60.2	.728
1959	71,008	126.70	60.2	.728
1960			60.2	.771
1961			60.2	.771
1962			60.2	.771
1963			60.2	.771
1964			60.2	.771

References

1. K. J. Arrow, H. B. Chenery, B. Minhas, and R. M. Solow, "Capital-Labor Substitution and Economic Efficiency," *Rev. Econ. Stat.*, Aug. 1961, *43*, 225–47.
2. M. J. Beckmann, "Einkommensvertei-

lung und Wachstum bei Nichtneutralem Technischen Fortschritt," *Jahrb. Nationalök. und Stat.*, Aug. 1965, *178*, 80–89.

3. P. A. DAVID AND T. VAN DE KLUNDERT, "Nonneutral Efficiency Growth and Substitution between Capital and Labor in the U.S. Economy, 1899–1960," *Am. Econ. Rev.*, June 1965, *55*, 357–94.
4. P. DIAMOND AND D. MCFADDEN, "Identification of the Elasticity of Substitution and the Bias of Technical Change: Impossibility Theorem," *Working Paper* 62, Institute of Business & Economic Research, University of California, Berkeley 1965.
5. E. D. DOMAR, "On the Measurement of Technological Change," *Econ. Jour.*, Dec. 1961, *71*, 709–29.
6. P. J. DHRYMES, "A Comparison of Productivity Behavior in Manufacturing and Service Industries," *Rev. Econ. Stat.*, Feb. 1963, *45*, 64–69.
7. D. W. JORGENSON AND Z. GRILICHES, "The Explanation of Productivity Change," *Rev. Econ. Stud.*, July 1967, *34*, 249–84.
8. J. W. KENDRICK AND R. SATO, "Factor Prices, Productivity and Economic Growth," *Am. Econ. Rev.*, Dec. 1963, *53*, 974–1004.
9. R. R. NELSON, "Aggregate Production Function and Medium-Range Growth Projections," *Am. Econ. Rev.*, Sept. 1964, *54*, 575–606.
10. R. SATO AND M. J. BECKMANN, "Neutral Inventions and Production Functions," *Rev. Econ. Stud.*, Jan. 1968, *35*, 57–67.
11. R. SATO, "Technical Progress and the Aggregate Production Function of Japan (1930–1960)," *Econ. Stud. Quart.*, March 1968, *19*, 15–24.
12. ———, "The Estimation of Biased Technical Progress and the Production Function," *Internat. Econ. Rev.*, forthcoming.
13. R. M. SOLOW, "Technical Change and the Aggregate Production Function," *Rev. Econ. Stat.*, Aug. 1957, *39*, 312–20.

Durability of Consumption Goods: Competition Versus Monopoly

By DAVID LEVHARI AND T. N. SRINIVASAN*

One encounters quite frequently assertions by laymen that firms in monopolized industries tend to shorten the durability of their products, compared with the durability of these goods under perfect competitive regimes. In the following we show conditions under which these assertions are valid.

It is well known that under the usual assumptions monopoly restricts output compared with the competitive case. If the costs of the competitive firms are just the same as the costs of plants of the monopoly, then the output of the monopoly will be smaller. In the competitive case, the assumption is that we have any number of potential firms which can produce output with the same cost function, and in the monopoly case the assumption is that we can build any number of plants with the same efficiency. In other words, in both cases we assume that the industry is operating under constant returns to scale.

However, unlike the textbook case in which, with given factor prices, costs are only a function of output level, here it is assumed that unit costs are a function of durability of the output. We assume the product to possess the property of sudden death, or one-hoss shay; we enjoy a constant flow of services from a unit of a durable good for a given time, and then the durable good evaporates at once. We assume a world where demand for services of durable goods is stationary, and firms continue to produce the same quantity of goods of constant durability each period.

We denote by p the price of a unit of service provided by the durable good. Further, if we assume that the units of product are measured in such a way that a unit of product continues to provide a unit of service at each point of time during its lifetime, then p is also the rental per unit of time of a unit of product. The demand function for services is $f(p)$. As is customary, we will assume $f'(p)<0$. If the durability in a certain stationary situation is N, and the amount of product produced each period is y, then the total services available at any point of time is Ny, and we have

$$(1) \qquad Ny = f(p).$$

If p is the rental rate, and r the rate of discount in the economy, then P_N, the price of the product with N years of life, should be

$$(2) \qquad P_N = p\frac{1 - e^{-rN}}{r}\cdot$$

The unit cost of production of a good with durability N is $c(N)$ with $c'(N)>0$. That is, the cost of producing y units of output with durability N is $yc(N)$.

In both the competitive and monopoly situation the volume of output and the durability of output have to be determined. Our main interest here is with the durability. In the competitive case, if we know the durability N and the price P_N we can, through (2), calculate the implied p and then using the demand schedule (1), find the volume of output y.

* The authors are, respectively, senior lecturer of economics at the Hebrew University, Jerusalem, and professor of economics at the Indian Statistical Institute, New Delhi. This paper was written while both authors spent a year at Stanford University.

The conditions for a long-run equilibrium of a competitive industry are that of zero profitability and equality of marginal cost with price. Given the assumption of constant returns to scale in production, these conditions reduce to

$$c(N) = P_N = p\,\frac{1 - e^{-rN}}{r}\,. \tag{3}$$

The condition for determination of N is that firms choose durability so as to maximize their profits. Since p is given to the single firm, the condition is

$$c'(N) = \frac{d}{dN}\,P_N = pe^{-rN}. \tag{4}$$

The second-order condition is that

$$\frac{d}{dN}\,[pe^{-rN} - c'(N)] < 0$$

or $-rpe^{-rN} - c''(N) < 0$, and using (4) the condition is

$$\frac{c''(N)}{c'(N)} > -r \tag{4a}$$

(the condition is automatically fulfilled in the case $c''>0$).

Dividing (4) by (3), we find

$$\frac{c'(N)}{c(N)} = \frac{r}{e^{rN} - 1}\,. \tag{5}$$

The solution of (5) is the N determined in the free competitive situation.

As one expects in this case, durability is related in a negative way to rate of discount r. If the rate r is lower, firms produce longer-lived assets.[1] Moreover, equilibrium durability depends only on the discount rate and the unit cost, and not on the demand function.

In the case where the industry is controlled by a monopoly, the firm would like to maximize its profits, i.e., to maximize $yP_N - yc(N)$ with respect to y and N. To get familiar textbook results, invert the demand equation (1) and obtain $p = g(Ny)$, or

$$P_N = \frac{1 - e^{-rN}}{r}\,g(Ny). \tag{6}$$

Profits R of monopoly in each period are:

$$R = y\,\frac{1 - e^{-rN}}{r}\,g(Ny) - yc(N).$$

Differentiating with respect to y, we obtain one of the required necessary conditions for a maximum:

$$\frac{\partial R}{\partial y} = \frac{1 - e^{-rN}}{r}\,g - c(N) + y\,\frac{1 - e^{-rN}}{r}\,Ng'(Ny) = 0,$$

---

[1] For this we prove that when

$$\frac{c'(N)}{c(N)} - \frac{r}{e^{rN} - 1} = 0 \tag{5}$$

is satisfied then $(dN/dr) < 0$. If

$$F(r, N) = \frac{c'(N)}{c(N)} - \frac{r}{e^{rN} - 1}$$

then $dN/dr = (-F_r/F_N)$. As

$$\frac{r}{e^{rN} - 1} = \frac{1}{N + rN^2 + r^2N^3 + \cdots}$$

it is obvious that $-F_r < 0$. Calculating F_N we find:

$$F_N = \frac{c''(N)}{c(N)} - \frac{[c'(N)]^2}{[c(N)]^2} + \frac{r^2e^{rN}}{(e^{rN} - 1)^2}$$

and using (5)

$$F_N = \frac{c''(N)}{c(N)} - \frac{r^2}{(e^{rN} - 1)^2} + \frac{r^2e^{rN}}{(e^{rN} - 1)^2} = \frac{c''(N)}{c(N)} + \frac{r^2}{e^{rN} - 1} = \frac{c'(N)}{c'(N)}\,\frac{c''(N)}{c(N)} + \frac{r^2}{e^{rN} - 1} = \frac{c''(N)}{c'(N)}\,\frac{r}{e^{rN} - 1} + \frac{r^2}{e^{rN} - 1}\,.$$

Using condition (4a) we finally obtain

$$F_N > -r\,\frac{r}{e^{rN} - 1} + \frac{r^2}{e^{rN} - 1} = 0.$$

As can be later observed this discussion is related to the discussion of the nature of the intersection of $c'(N)/c(N)$ with $r/e^{rN} - 1$ and as a matter of fact the same conclusion can be shown diagrammatically.

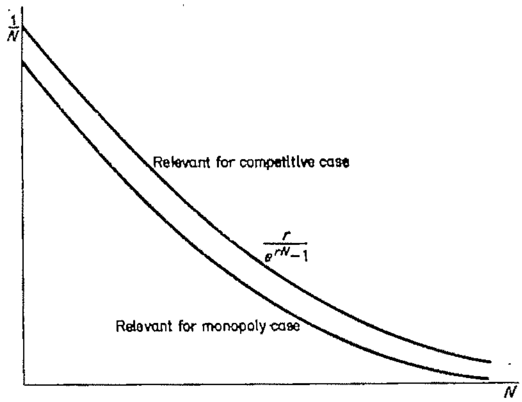


FIGURE 1

or

$$(7)\quad \begin{aligned} c(N) &= P_N\left(1 + Ny\frac{g'}{g}\right) \\ &= P_N\left[1 - \frac{1}{\eta}\right], \end{aligned}$$

where η is the elasticity of demand for services of the durable good. This is the usual condition for optimum output of a monopoly (the only relevant region for a monopolist is where $\eta > 1$; otherwise the monopolist can increase profits by reducing his output). Differentiation of R with respect to N yields

$$\frac{\partial R}{\partial N} = y\left[e^{-rN}g(Ny) + y\left\{\frac{1 - e^{-rN}}{r}\right\} \cdot g'(Ny) - c'(N)\right] = 0,$$

or the other necessary condition for a maximum is:

$$(8)\ c'(N) = g\frac{1 - e^{-rN}}{r}\left[\frac{r}{e^{rN} - 1} + y\frac{g'}{g}\right]$$

or

$$(9)\quad c'(N) = P_N\left[\frac{r}{e^{rN} - 1} - \frac{1}{N\eta}\right].$$

Division of (9) by (7) yields

$$(10)\quad \frac{c'(N)}{c(N)} = \left[\frac{\dfrac{r}{e^{rN} - 1} - \dfrac{1}{N\eta}}{1 - \dfrac{1}{\eta}}\right].$$

Comparing (10) with (5), we observe that on the left-hand side we have the same function of N, $c'(N)/c(N)$. The right-hand side, on the other hand, is always smaller in (10) than in (5). Thus,

$$\begin{aligned} &\frac{\dfrac{r}{e^{rN} - 1} - \dfrac{1}{N\eta}}{1 - \dfrac{1}{\eta}} - \frac{r}{e^{rN} - 1} \\ &= \frac{\dfrac{r}{e^{rN} - 1}\left[1 - \left(1 - \dfrac{1}{\eta}\right)\right] - \dfrac{1}{N\eta}}{1 - \dfrac{1}{\eta}} \\ &= \frac{1}{\eta - 1}\left[\frac{rN + 1 - e^{rN}}{N(e^{rN} - 1)}\right] < 0 \end{aligned}$$

for $\eta > 1$, $N > 0$.

The right-hand side of (5) is a declining function of N so that the situation in two cases can be depicted by Figure 1:

Now the question whether durability N is longer in the competitive case, compared with the monopoly case, depends on the way $c'(N)/c(N)$ crosses these two curves. That is, if $c'(N)/c(N)$ crosses these curves from below and we have just one solution, then the competitive market gives longer durability. For the competitive solution,

$$\frac{\partial}{\partial N}\left[\frac{c'(N)}{c(N)} - \frac{r}{e^{rN} - 1}\right] = \frac{c''c - (c')^2}{c^2} + \frac{r^2e^{rN}}{(e^{rN} - 1)^2}.$$

We can show as follows that at the intersection point the right-hand side is positive, and at this point can be written

$$\frac{c''c + (c')^2(e^{rN} - 1)}{c^2},$$

and for this to be positive we need

$$\frac{c''c}{(c')^2} + e^{rN} - 1 > 0$$

or

$$\frac{c''}{c'}\,\frac{e^{rN} - 1}{r} + e^{rN} - 1 > 0.$$

Now remembering the second-order condition

$$\frac{c''}{c'} > -r,$$

$$\text{(4a)}\quad \frac{c''}{c'}\,\frac{e^{rN} - 1}{r} + e^{rN} - 1 > -r\,\frac{e^{rN} - 1}{r} + e^{rN} - 1 = 0.$$

That is, c'/c intersects always from below in this case.[2] Thus, for a maximum, the situation in this case is the one depicted in Figure 2:

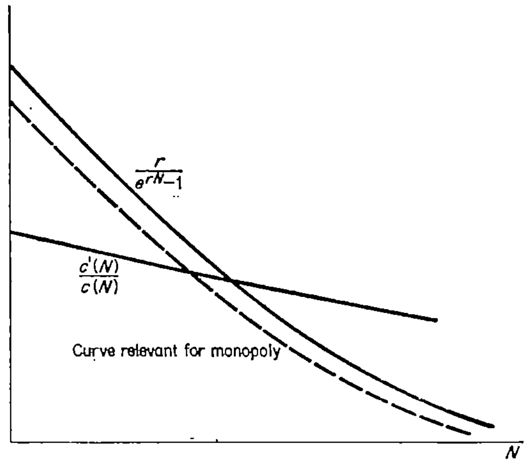


FIGURE 2

In the monopoly case, the decisions on y and N are interdependent and the second-order condition does not yield a simple rule on the relevant slopes. However, if the situation is the one in Figure 2, in which in the monopoly case as well as in the competitive case the intersection of c'/c is from below, we obviously find that monopolization restricts durability of its product.

In the case where $c'(N)/c(N)$ is rising, we always find that durability in the monopoly case is shorter.

We may now give some simple examples. Assume that the unit cost, $c(N)$, is of the power function type $c(N) = AN^{\alpha}$, with $\alpha < 1$. (This is needed since, as we noted earlier, at the intersection point the elasticity of the cost function must be below unity.) The demand for services is $Ny = Bp^{-\beta}$ and $\eta = \beta$. In this case,

$$\frac{c'(N)}{c(N)} = \frac{\alpha}{N}$$

and condition (5) for durability in the competitive case is

$$\frac{\alpha}{N} = \frac{r}{e^{rN} - 1},$$

or

$$\text{(11)}\qquad N = \alpha\,\frac{e^{rN} - 1}{r}.$$

The left-hand side is the 45° line, while the right-hand side possesses a positive second derivative; hence, the intersection is unique and in the right direction for a maximum.

In the monopoly case, the condition for optimal durability is

[2] Equation (5) can be written as

$$\frac{Nc'(N)}{c(N)} = \frac{rN}{e^{rN} - 1};$$

that is, at the point of optimal durability the elasticity of the cost function should be $rN/e^{rN}-1$. The function on the right-hand side is a decreasing function of N which assumes the value 1 at $N=0$. Therefore at the optimum point N we require the elasticity of the cost function to be smaller than 1.

$$\frac{\alpha}{N} = \frac{\dfrac{r}{e^{rN} - 1} - \dfrac{1}{N\beta}}{1 - \dfrac{1}{\beta}},$$

or

$$\frac{\alpha(\beta - 1) + 1}{\beta - 1} \frac{1}{N} = \frac{\beta}{\beta - 1} \frac{r}{e^{rN} - 1},$$

and finally

$$N = \frac{\alpha(\beta - 1) + 1}{\beta} \frac{e^{rN} - 1}{r}. \tag{12}$$

Since

$$\alpha < 1, \frac{\alpha(\beta - 1) + 1}{\beta} > \alpha, \tag{13}$$

and we find that N is smaller for this case (Figure 3).

Another very simple example is the one in which the cost function is an exponential function of N, say $Me^{\gamma N}$, in which case $c'/c = \gamma$ and, regardless of demand conditions, monopoly produces shorter-lived consumer goods. If the demand function is assumed to be that in the previous example, then equation (5) assumes the form:

$$\gamma = \frac{r}{e^{rN} - 1},$$

or

$$N = \frac{\log\left(1 + \dfrac{r}{\gamma}\right)}{r}. \tag{14}$$

In the monopoly case,

$$\gamma = \frac{\dfrac{r}{e^{rN} - 1} - \dfrac{1}{N\beta}}{1 - \dfrac{1}{\beta}}, \tag{15}$$

or

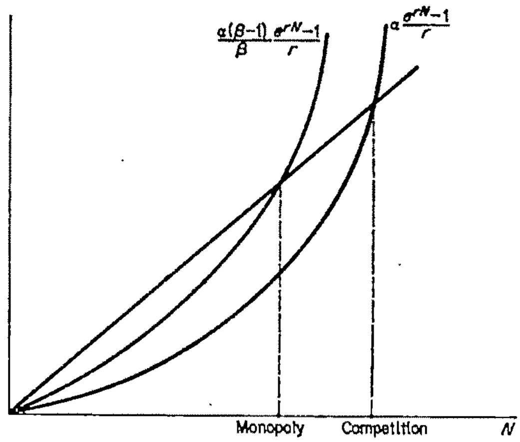


FIGURE 3

$$1 + (\beta - 1)\gamma N = \beta \frac{rN}{e^{rN} - 1}. \tag{16}$$

It is again easy to see that if $\beta > 1$, a solution always exists (Figure 4) for the monopoly case as well, and the good is less durable than under competition.

One policy question that may be asked is whether public control of durability will benefit the consumer.[3] That is, if the government sets standards for durability while the monopoly controls output, what

[3] We thank Professor Kenneth Arrow for raising this question in a private conversation.

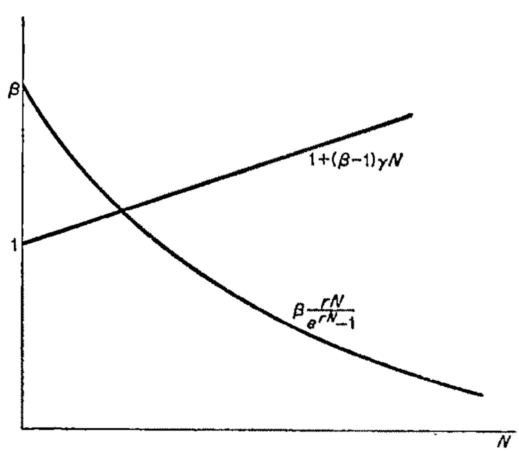


FIGURE 4

is the expected impact on consumers? In the following section we prove that as long as the standards are effective in the sense that N set by the government is longer than the one chosen by the monopolist, but not longer than the competitive one, consumers are better off; it means that consumers enjoy a larger volume of services and, of course, the rental price of services is lower.

Ny is the volume of services supplied, denoted by z. The monopolist wishes to maximize his profit:

$$R = \frac{z}{N} \frac{1 - e^{-rN}}{r} g(z) - \frac{z}{N} c(N)$$

subject to the N given by the government. By differentiation with respect to z, we obtain the maximum profit condition:

$$(17) \quad g(z) \frac{1-e^{-rN}}{r}\left(1-\frac{1}{\eta(z)}\right)-c(N)=0$$

which is merely rewriting (7) in terms of z.

To prove the assertion made, we have to show that for a larger N the monopolist will choose a larger z. That is, using (17), we show that $dz/dN>0$.

Denote the left-hand side of (17) by $\psi(z, N)$, then: $dz/dN = -(\psi_N/\psi_z.)$ We easily find:

$$(18) \qquad \psi_z = \frac{1}{N} \frac{\partial R}{\partial z^2} < 0$$

as the second order condition for the maximization of R with respect to z. Differentiation of ψ with respect to N yields:

$$(19) \quad \psi_N = g(z)e^{-rN}\left(1 - \frac{1}{\eta(z)}\right) - c'(N).$$

Using (17) and (19) this can be rewritten as:

$$(20) \qquad \psi_N = c(N) \frac{re^{-rN}}{1 - e^{-rN}} - c'(N)$$

or

$$(21) \qquad \frac{\psi_N}{c(N)} = \frac{r}{e^{rN} - 1} - \frac{c'(N)}{c(N)}.$$

Assume now that the standard determined by the government is such that N is not larger than the competitive N; then using the nature of the intersection of $c'(N)/c(N)$ and $r/(e^{rN}-1)$ (Figure 2), we find that

$$\frac{\psi_N}{c(N)} > 0 \quad \text{or } \psi_N > 0.$$

Hence, under the mentioned conditions,

$$\frac{dz}{dN} = -\frac{\psi_N}{\psi_z} > 0$$

as originally stated.

All our discussions have been stated in terms of durability, but one suspects that the same results will hold for other (not as easily measured) concepts, such as quality. That is, besides restricting output, the monopolist will produce goods of lower quality.

Customs Union and The Theory of Tariffs

By Sven W. Arndt*

The standard analytical vehicle of customs union theory consists essentially of a two-commodity, three-country model of the trading world. Within this framework, two of the three countries form the union, while the third "country" comprises the "rest of the world." The latter is either assumed to contain a single country or, where several economies are involved, these are taken to be perfectly homogeneous, so that differentiation among them is immaterial to the analysis.[1]

Many significant results have been obtained with the aid of this apparatus, which has in spite, or perhaps because, of its basic simplicity taken the analysis of customs union a long way since the problem was first formalized by Viner [18]. Nevertheless, the model does suffer from the burden of several major constraints. Not the least of these involves the treatment of the non-union or outside word as a single unit. In the present paper, we shall develop a model that will permit us to break up the rest of the word into several heterogeneous components. The smallest version of this expanded model will contain four countries—two within the union and two without—but the number of countries may be readily increased.

Among the principal results of this extension, the traditional conclusion of customs union theory that terms-of-trade movements will lead to deterioration in the welfare of the "outside world" will be shown to be of limited relevance to the multiple-country situation. It will be shown further that when exclusion from the union involves more than a single country, the incentives to elect nonmembership will be increased for some countries by the final form of the union. An attempt will also be made to demonstrate the superiority of customs union as a particular type of tariff policy.

The model will be introduced in Section I; it will be used there to examine the case of the large union which imposes a prohibitive tariff. This will bring out the importance of movements in the terms of trade in determining the welfare effects of customs union. In Section II, the more general case of the nonprohibitive tariff will be taken up. Section III will deal with the existence of tariff policies as alternatives to customs union. A final summary concludes the paper.

I. *Large Union: Prohibitive Tariff*

The initial purpose of this section is to set out the framework of the analysis. Among the basic properties of the model is the assumption that full employment is maintained throughout and that factor supplies are fixed. Further, perfect competition is assumed to prevail. Commodities are homogeneous and mobile between countries without transportation

* The author is assistant professor of economics at the University of California at Los Angeles. He is indebted to William R. Allen for his valuable comments on an earlier draft. He is, of course, solely responsible for any remaining errors. Fellowship support from UCLA's Committee on International and Comparative Studies with funding from the Ford Foundation is gratefully acknowledged.

[1] See, for example, the papers by Lipsey [10] and Michaely [13]. Vanek [16] uses a similar framework, but he does attempt some generalizations beyond the three-country world. Cooper and Massell [4], on the other hand, conduct their analysis entirely in terms of a single-commodity, three-country model.

costs. Factors of production are perfectly immobile internationally. Tariffs are imposed for regulatory purposes only; all proceeds are returned to consumers as income subsidies. Individual tastes and preferences are such as to justify the use of community indifference curves.[2] Finally, all production transformation functions are assumed to be concave to the origin.

Suppose, then, that the world may be divided into two groups of countries producing commodities X and Y. Group W_A is the set of all actual or latent exporters of Y, while group W_B exports X. The groups are necessarily defined with reference to the prevailing terms of trade, for it is clearly possible that alterations in the international price ratio will introduce trade reversals which will cause countries to shift between groups.

An essential feature of the present model is its use of aggregate reciprocal demand functions. Such functions are constructed by taking the radial sum at various terms of trade of the offers of the members of a given group.[3] The shape of the aggregate curve is generally dependent upon the individual component curves; its smoothness increases with the number of countries and with the similarity among the underlying offer curves; its elasticity is determined by the elasticities of the individual curves. Where individual countries are small, changes in the behavior of a single member of a group will leave unaffected the group reciprocal demand curve and thus the international terms of trade. A large country, on the other hand, will cause shifts in its offer curve to be reflected in the aggregate curve.[4]

Group W_A, the exporter of Y, may consist of one or more countries. We assume for simplicity, but without loss of generality, that it contains two countries, A_1 and A_2. Suppose further that A_1 is the lowest-cost producer of Y, so that it will be a supplier of Y and an importer of X at lower relative prices of Y than any other country. Specifically, if OA_1' and OA_2' are the countries' relevant offer curves, A_1 will be the sole exporter of Y at all terms of trade equal to or steeper than those given by the ray through point W (the line not being drawn to avoid cluttering up the diagram). Consequently, the group offer curve will be coincidental with A_1's offer curve up to point W, i.e., country A_2's curve is tangent to the line OW at the origin. When the price of X is high enough to cause country A_2 to become an exporter of that commodity, i.e., to produce a negative offer of Y, the country's demand

[2] This assumption is of special importance in view of the subsequent use of reciprocal demand curves. It is assumed that the distribution of income is held constant by the authorities. This is clearly a restrictive assumption which ignores, among other things, the possibility raised by Johnson [7] and Metzler [12] that tariff-induced changes in the terms of trade will alter the distribution of income. Its virtue lies in the fact that it eliminates the problem that tariff alterations and formation of customs union may cause the reciprocal demand curves of a given country to cross. For more detailed discussions of community indifference curves, see Michaely [13] and Vanek [17, Ch. 16].

[3] Aggregate reciprocal demand curves were implied by Graham [6] and have been derived by Becker [2] and Elliot [5].

[4] We may note the special case in which the members of a group possess identical offer curves. The resulting aggregate curve will be perfectly smooth and will possess the convenient property that the entry or withdrawal of one or more of these countries will leave the elasticity of the aggregate curve invariant with respect to a given price ratio. Kemp [9] has discussed this invariance in connection with various types of demand shifts. While he was concerned primarily with individual country offer curves, the present application is a natural extension of his analysis.

When offer curves differ among countries, the elasticity of the group offer curve will be determined by the relative propensities to consume importables in the component countries and by the individual shares in total group imports, on the one hand, and by a compensated price elasticity for the group, on the other. Both income and price effects will essentially be weighted averages of the individual component country terms. A disaggregated elasticity of demand has been derived for a single country composed of two individuals by Bhagwati and Johnson [3]. That analysis is readily adaptable to the present situation.

for the latter commodity must be added to the aggregate demand of group W_B.

At terms of trade flatter than OW, country A_2 begins to export commodity Y. The group reciprocal demand curve, OWW_A', reflects this increased supply of Y on world markets by moving away from OA_1' at this point. The intersection of OWW_A' and OW_B', which is group W_B's aggregate offer curve and which may be derived analogously, determines the world terms of trade. If the country curves had been free-trade curves, the resulting terms of trade would give the free-trade world price ratio as well as the internal relative prices in the various countries. If, on the other hand, OA_1', OA_2', OB_1' and OB_2' are all revenue-redistributed offer curves corresponding to various individual tariffs imposed by the four countries, then OWW_A' and OW_B' are the resulting group offer curves and OT_t measures the terms of trade of a tariff-ridden world. This latter case is in fact the one depicted in the first quadrant of Figure 1. Free-trade offer curves have, in general, not been drawn in order to keep the diagrams as simple as possible.

Let us assume, therefore, that quadrant 1 represents the situation as it exists prior to formation of the customs union. World-trade equilibrium is located at P, the point of intersection of the revenue-redistributed group offer curves. The terms of trade are given by OT_t. In this situation, different cost, demand, and tariff conditions will prevail within the various countries. To illustrate, the revenue-redistributed offer curve (OA_2') of country A_2 corresponds to an ad valorem rate of tariff of OM/MN. The internal price ratio in A_2 is indicated by RM. Point R represents equilibrium for that country. As Johnson [8, Ch. 2] has shown, the terms-of-trade line cuts the indifference curve which passes through R, whereas the internal price line is tangent to that indifference curve. In the free-trade case, on the other hand, where internal and international price ratios are identical, the indifference curve passing through the equilibrium point will be touched by the common price line.

We may think of country A_1 as having imposed a higher ad valorem rate of duty, given by OK/KL and producing an internal price ratio of SK. At *constant* terms of trade, the internal price of X will be higher in A_1 than in A_2, given the higher rate of taxation in A_1. When countries are large enough to affect the terms of trade, however, differences in size will tend to offset (or augment) differential effects of varying tariff rates on prices.[5]

Suppose now that countries A_2 and B_2 form a customs union with prohibitive tariff. Countries A_1 and B_1 thus constitute the "outside world." In the latter, trade equilibrium will shift to P', the point of intersection of the revenue-redistributed offer curves of the truncated groups, provided that neither group reacts to the union by altering its tariff policies. The new outside terms of trade are OT_w; they represent a worsened trade situation for the country importing commodity Y.

In the third quadrant, OA_2' and OB_2' are drawn along with the countries' free-trade offer curves, OA_2 and OB_2. With a prohibitive external tariff the intra-union terms of trade will be OT_u. It would thus appear that the terms of trade facing A_2 and those facing its former group, and for

[5] The word "size" is used here in the broad sense, encompassing both demand and supply conditions. It is well known that a tariff will leave the internal price ratio unchanged if the sum of the marginal propensity to import in the taxing country and the foreign elasticity of demand for that country's exports equal unity. See, for example, Meade [11, p. 74]. This condition, however, holds strictly only for the case in which the given country is the sole supplier of the commodity; in the presence of competing suppliers the conditions must be amended to include the effects on third country offers of the given country's tariff policy.

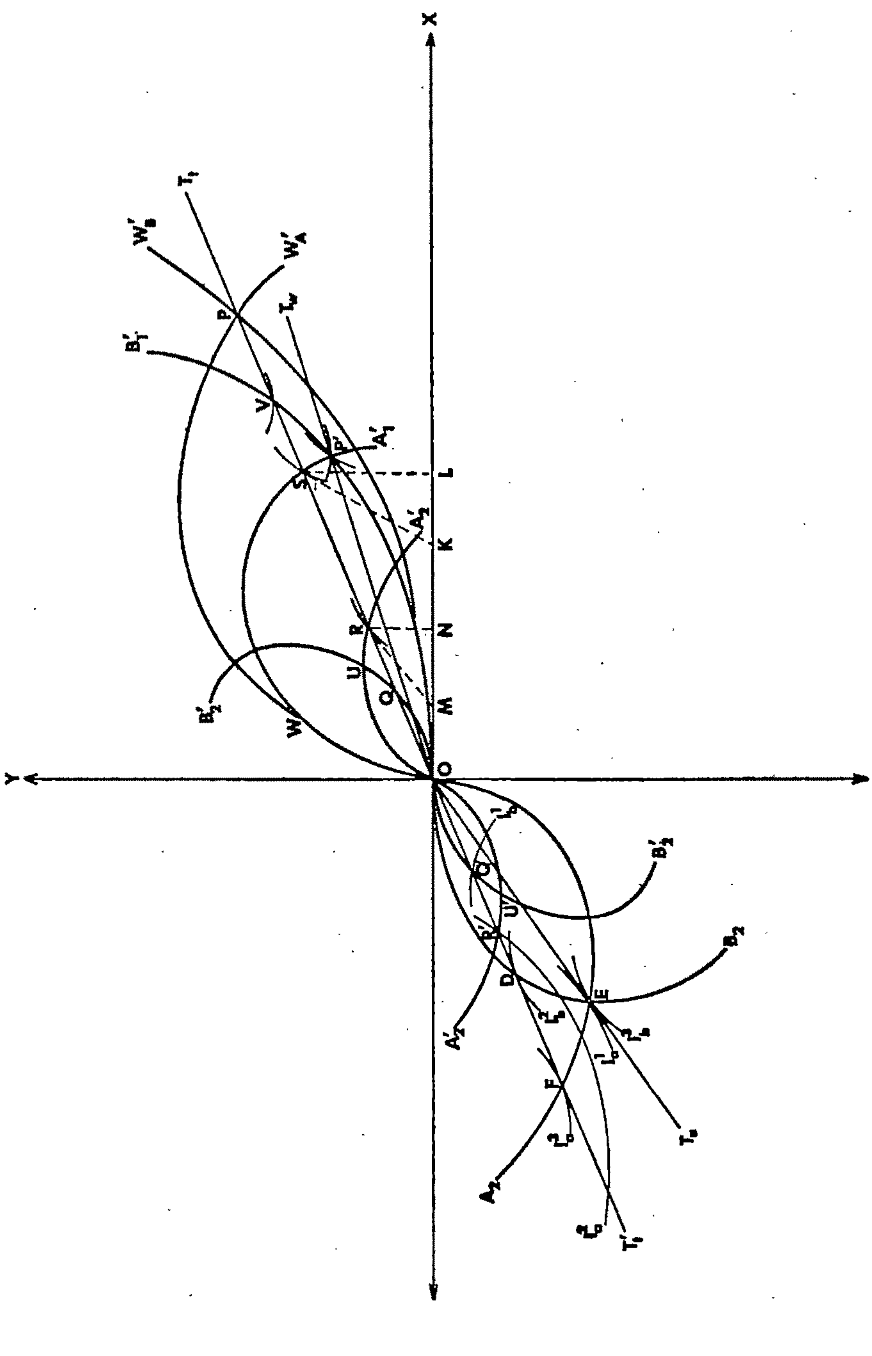


FIGURE 1

symmetrical reasons B_2 and its former group, have moved in opposite directions. This, however, need not always occur. It is possible that, for an appropriate choice of countries and, within these, an appropriate choice of ad valorem tariffs, the final intra-union terms of trade will move in the same direction as those in the outside world. As a special case, it is possible for the terms of trade in either the union or the outside to remain unaltered at OT_t while those in the opposite block take on a new value.

Finally, there is one case, namely, union of "similar" economies,[6] in which we can make unequivocal statements about the direction of the terms-of-trade shifts. For example, if two of a large number of countries in W_A form a union, the terms of trade outside the union will clearly move in favor of group W_A, and thus the intra-union price ratio *must* move in the opposite direction. In the absence of the latter increase in the relative price of X, the commodity will not be produced within the union.[7]

A significant aspect of customs union theory has been its attempt to predict the direction of terms-of-trade changes. Whether the terms of trade move in favor or against the "inside" or the "outside" is of considerable import not only to the respective union partners, but to those who remain nonmembers as well.[8] As the foregoing discussion makes clear, however, it is generally undesirable, if not meaningless, to deal with the "outside world" as a single entity when that world contains several distinct members. We can see in our present case of the prohibitive tariff, that every movement of the terms of trade will be favorable to some countries and adverse to others. There is only one clear case and that involves union of similar economies in which *all* exporters of one commodity, say, Y, join the union. Then, indeed, will the relative price of X rise in the union and fall in the outside world, producing trade reversals in some countries and leaving both blocks with uniformly worsened terms of trade.

Country A_2, having moved from R $(=R')$ to E, is worse off after customs union. The total change in welfare may be separated into two components. The first, which tends to increase the country's welfare, is the result of tariff elimination with its increased specialization in production and its attendant readjustments of consumption. It is equivalent to a movement from R' and indifference curve I_a^2 to F and curve I_a^3 at *unchanged* terms of trade. Removal of the tariff produces a reallocation of resources in favor of Y and substitution in consumption toward X by changing the internal price ratio from MR to OT_t'. The second component, which tends to reduce A_2's welfare, arises from changes in the terms of trade from OT_t' $(=OT_t)$ to OT_u. It may be shown in perfectly analogous fashion that both component movements are favorable for B_2, whose welfare is increased unequivocally. By joining the union, the country gains both because its volume of trade increases and because its terms of trade improve. In order to determine the net change in union welfare, intercountry comparisons will become necessary, except for the possibility that the gainer may be able to compensate the loser and still prefer the union.

The foregoing analysis may be applied to the changes which will occur in the outside world. The movement from S to P' constitutes a welfare improvement for A_1: its terms of trade improve and the

[6] In Vanek's terminology [16], union of countries from the same group constitutes union of similar economies. The opposite situation produces union of dissimilar economies.

[7] It remains to be seen in a later section to what extent relaxation of the prohibitive tariff alters this conclusion.

[8] Cf., for example, Mundell [15] and Vanek [16].

internal relative price of imports falls. The welfare of B_1, on the other hand, deteriorates as a result of the formation of the union. The positions of the two countries are affected only by terms-of-trade shifts, given the unaltered degrees of their trade restriction.[9] It can be seen that the attitudes of A_1 and B_1 toward formation of the union will be very different. B_1 will try to oppose the union, fearing a worsening in its terms of trade. Country A_1 will support the union, but should clearly refrain from joining it, for it has much to gain from its increased market power in the outside world.

II. *Large Union: Nonprohibitive Tariff*

When the union tariff is less than prohibitive, OT_u and OT_w in Figure 1 no longer represent the final terms of trade. As the tariff is lowered to nonprohibitive levels, substitution in production within the union toward the product whose price is rising, namely Y, and substitution in consumption toward the product whose price is falling, namely the import commodity X, will tend to push OT_u in a clockwise direction, while analogous adjustments in the outside world will rotate OT_w in a counterclockwise manner. These rotations, however, will stop short of bringing the internal price vectors in the two blocks into equality, unless both remove their respective external tariffs completely.

If we assume that all tariff changes which actually occur are those involving the union, then OA_1' and OB_1' will continue to represent the outside world's ruling offer curves. The union offer curves, on the other hand, will be altered as the external union tariff is reduced, producing nonzero excess demands within the union for the two goods at their new prices. For the extreme case of a zero tariff, OA_2 and OB_2 will be the relevant offer curves; in this case, the final terms of trade will be steeper than OT_t, given the relative magnitudes of the pre-union tariffs of the partners and the relationships between curves OA_2' and OA_2 and OB_2' and OB_2.

In the general case, the world terms of trade will fall somewhere between the latter ratio and OT_w, the exact location being determined by the final size of the union tariff. It is apparent that the two union partners will be induced to pursue conflicting objectives with respect to the

[9] The present case raises the following question: will the tariff levels in A_1 and B_1, which were established in accordance with the pre-union trading situation, be consistent with post-union conditions? Suppose that prior to the union country A_i is relatively small within group W_A and that group W_B's offer curve is highly elastic. A_i's ability to alter the terms of trade will be small and its optimum tariff low. Suppose, next, that several members of the two groups join a prohibitive customs union. This increases A_i's power within the newly constituted group W_A. Assuming that the reduction in the size of W_B has raised the curvature of its reciprocal demand curve, it follows that optimum strategy, especially in the absence of retaliation, may now lead A_i to increase its tariff; and if this is generally applicable to the countries on the outside, the average level of protection will increase.

But is this the general case; or is it possible for optimum tariff policy to require tariff cuts? Suppose that all members of W_A and W_B enter a prohibitive union except for countries A_1 and B_1. Quadrant 3 then shows the situation in the outside world after formation of the union. At unchanged pre-union tariffs, the terms of trade are given by the ray (not drawn) through U'. Through that point will pass indifference curves for each country which will be touched by their respective internal price lines. It has been shown by Johnson [8, Ch. 2] and others, that if at U' the offer curve of the passive country is tangent to the internal price line of the tariff-imposing country, the latter country cannot gain by altering its tariff. If, on the other hand, the offer curve *cuts* the price line of the taxing country from the side of the line which lies between that line and the taxing country's export axis, the tariff should be increased. In the reverse case, the tariff will have to be reduced since it will exceed its optimum level.

The point of intersection, U', will be determined by the rates of duty established in the two countries prior to formation of the union. These rates will be affected by the sizes of the two countries within their respective groups, by the elasticities of the various offer curves and by the past history of tariff competition. There is no presumption, therefore, that optimal tariff strategy for the countries on the outside will always lead to increases in the rate of protection.

optimum level of the common tariff. Country B_2's gain from union will be smaller, the lower is that tariff; that country will thus tend to resist reductions in the common tariff. A_2, on the other hand, would see its earlier losses reduced by lower external tariffs, and is thus likely to pursue trade liberalization. Resolution of the conflict will probably involve redistribution of union tariff revenues to the residents of country A_2 plus compensating transfer payments from B_2 to A_2. In the case at hand, we see that the potential for such transfer payments exists.[10]

III. *Customs Union and Alternative Tariff Policies*

The analysis of customs union has traditionally concerned itself with the conditions under which countries will find regional or group tariff discrimination preferable to unilateral tariff policy. Alternatively stated, the theory has attempted to define the range of situations in which regional free trade will be preferable from various points of view to generally restricted trade. In the development of the theory, however, it has been suggested that some members of a proposed union would be at least as well off, and quite possibly better off, if they unilaterally and nonpreferentially reduced their own tariffs. This point is implicit in Lipsey [10] and Michaely [13], and it constitutes the major argument of Cooper and Massell [4].[11] In all three papers, terms-of-trade effects are assumed absent.[12] Hence, for each country acting on its own, the welfare-maximizing tariff is the zero tariff. For any given country, a small customs union may or may not be as beneficial as unilateral freeing of trade, depending upon whether the lowest-cost producer of its importable commodity becomes a member of the union or remains outside.[13] What remains to be established is whether these conclusions are applicable to the case of the union which is large enough to alter the terms of trade.

Consider Figure 2. Let $OW_A{}^o$ and $OW_B{}^o$ represent the group offer curves consistent with various tariff levels in various countries, but with no trade restriction in countries A and B.[14] For the latter two countries, the relevant offer curves are OA^o and OB^o, respectively. The world terms of trade are given by OT_o; A is in equilibrium at point $Q_a{}^o$ on indifference curve $I_a{}^o$, while B is at $Q_b{}^o$ on $I_b{}^o$. It is possible for either country to raise its welfare by imposing an *appropriate* tariff, particularly if neither the other country nor the countries making up the rest of the world retaliate. Thus, if country A alone imposes a tariff such that OA' becomes its revenue-redistributed offer curve, $OW_A{}'$ constitutes the new A-group reciprocal demand curve and OP_a gives the new world terms

[10] When its external tariff is less than prohibitive, the union will collect tariff revenues, which may be shared by the residents of the union in a variety of ways. Since the welfare of the member countries will be affected by the particular distribution scheme adopted, the willingness of a given country to participate in the union will be influenced, among other things, by its share in union tariff revenues.

As for the transfer payments from one member country to another, it has been shown by Meade [11], for example, that the introduction of such payments will produce shifts in offer curves due to the fact that national expenditure will no longer equal national income.

[11] For an argument supporting customs union as a form of coalition see Arndt [1].

[12] Vanek [16] is an exception as far as this assumption is concerned.

[13] We are presently concerned solely with the specialization and terms-of-trade arguments of customs union. It is recognized that whatever the benefits or losses associated with these factors, there may be numerous other reasons which may favor a particular customs union.

[14] The curvature of the group offer curve has been accentuated deliberately in order to clarify the exposition of the argument. Use of more elastic aggregate offer curves would produce crowding of the various points in the diagram, since the relative terms-of-trade shifts upon which much of the analysis is based would be correspondingly smaller.

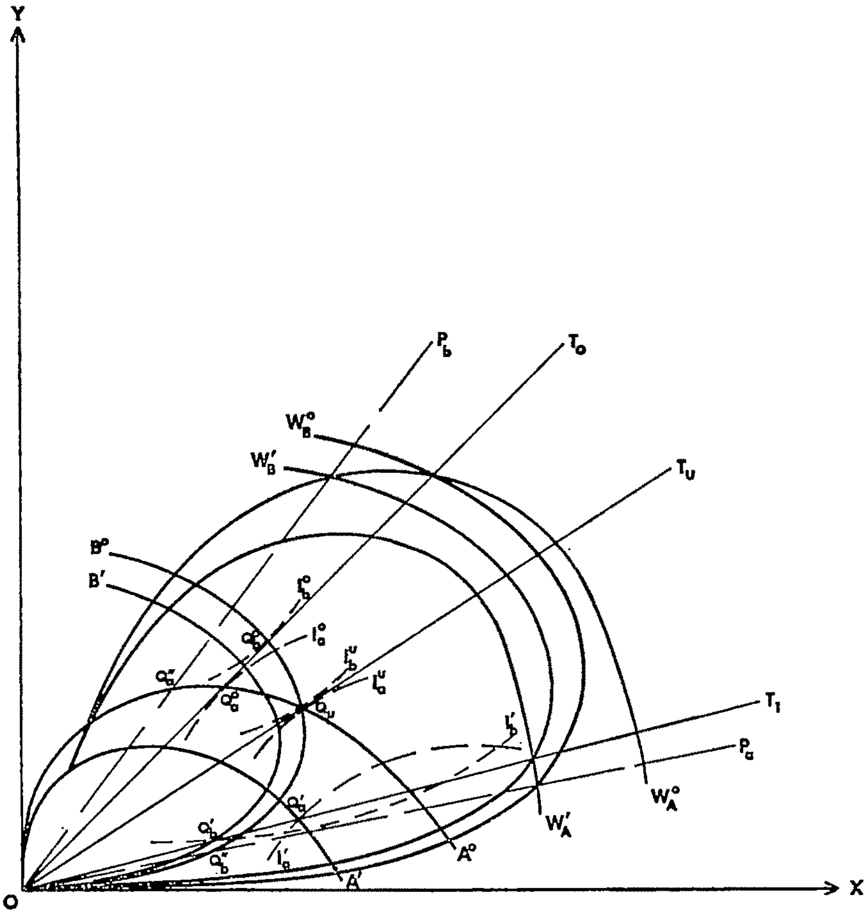


FIGURE 2

of trade. Country A's new equilibrium will be established at the point of intersection of OP_a and OA'; the welfare gain implicit in this policy is clear at once. Analogous remarks apply to country B and line OP_b. Retaliation by other countries, however, is likely to occur and hence to reduce or offset the favorable effects of unilateral tariff manipulation. This point can be ignored in the small union case, but it is of crucial importance in the present situation.

Assume next that both countries have in fact pursued past tariff policies and that OA' and OB' indicate their respective revenue-redistributed offer curves. OW_A' and OW_B' are the resulting group offer curves. The terms of trade are given by OT_1; A is now at Q_a' on indifference curve I_a' and B is at O_b' on I_b' A's welfare has improved relative to its no-tariff situation, while that of B has worsened. Starting with these conditions, and all other things unchanged, neither country's preferred strategy consists of unilateral, nonpreferential tariff *abolition*, since this would lead to inferior positions Q_a'' and Q_b'', respectively.

However, it may be equally impossible for either country to gain from further tariff manipulation. An increase in country A's tariff at point Q_a', for example, will shift the offer curves belonging to A and its group inward and move the terms of

trade, OT_1, in a clockwise direction. The proportional displacement of the terms of trade will be smaller relative to the given proportional increase in the tariff, the smaller is the country within its own group and the larger is the number of trade reversals produced by the movements in the terms of trade. Trade reversals will inhibit the change in the terms of trade by placing pressure on OW_B' to move inward and OW_a' to move outward. In the limit, a tariff which causes OA' to shift while it leaves OT_1 unchanged will clearly reduce A_1's welfare.

As an alternative, suppose that the two countries form a union, initially with a prohibitive tariff. The union price ratio is then given by OT_u and trade equilibrium within the union is established at Q_u with A on indifference curve $I_a{}^u$ and B on $I_b{}^u$. Country B's welfare has improved relative to the initial tariff situation, while A's has worsened. It is possible, however, for appropriate transfers from B to A to enable the latter to reach indifference curve I_a', thus making it indifferent between the union and the earlier tariff situation, while leaving union membership preferable to B. A marginally larger transfer payment will make customs union more desirable for both countries.[15]

The foregoing conclusion depends upon several assumptions. First, the pre-union tariffs were assumed optimum tariffs for the two countries. Second, it was noted in an earlier section that the particular elasticities of the various curves, as well as the relationships between individual countries and their groups will affect the final outcome; for example, our results will change if elasticities are such that the indifference curves associated with the pre-union tariffs intersect in the area below curve OA^o. Third, and most important for present purposes, the union tariff was prohibitive.

In order to examine the nonprohibitive case, we note the following. The price ratio in the outside world which results from a prohibitive union tariff may coincide with or it may lie on either side of OT_1, and this set of ratios clearly includes those which lie above OT_u. The actual relative position will be determined both by the sizes of the two countries and by their respective pre-union tariffs.[16] Assuming that union autarky produces terms of trade in the outside world which are flatter than OT_u, the intra-union price of Y will increase as the tariff is lowered; at unchanged outside tariffs, the price of Y will fall in the outside world. From the union's point of view, this has two important results. First, the lower tariff, and thus the more advantageous price of Y, will make the union increasingly more desirable to country A without eliminating its attractiveness to B, unless the final price ratio within B is equal to or worse than OT_1. Second, the tariff will produce revenue which the union can distribute as compensation to A. It would thus appear, that country A, having pursued an aggressive tariff policy prior to union, will display considerable interest in a moderate union tariff. The argument for the case in which the world terms of trade under union autarky lie above OT_u may be derived in analogous fashion. We see in this instance that A will push for high union tariffs.

We have assumed throughout that the rest of the world remains completely passive. Yet it is clear that the changes we have discussed will affect other countries. For instance, unilateral tariff re-

[15] As indicated in footnote 10, introduction of transfer payments will relocate offer curves and their origins.

[16] We recall that OW_A' and OW_B' correspond to the given pre-union, nonprohibitive tariffs in A and B, respectively. Complete withdrawal by the two countries from trade with the rest of the world will produce outside offer curves which lie inside OW_A' and OW_B'.

duction by A will worsen the terms of trade of others in group W_A: these countries may thus initiate retaliatory tariff policies. Similarly, in the post-union world the countries in one of the groups will be worse off in the sense that the terms of trade will have moved against them; they too may retaliate. The union is thus likely to produce alterations in the tariffs of many countries, and, as shown earlier, not all of these changes will result in higher rates of protection.

IV. *Conclusion*

Whenever a customs union assumes proportions which provide it with sufficient economic power to influence the prevailing terms of trade, the resulting impact on welfare may either augment or offset the welfare changes due to preferential tariff reductions. Some of these possibilities have been investigated, particularly by Vanek [16], within the framework of a three-country model.[17] In the preceding sections we have attempted to widen the inquiry by moving away from the customary three-country analysis. Certain conclusions have emerged from this broadened investigation. First, it is no longer particularly meaningful—except for union among all of the exporters of a given commodity—to argue that the customs union will do one thing or another to the terms of trade faced by the outside world. Rather, shifts in the terms of trade will produce welfare gains for some countries and losses for others. Whether the net effect constitutes loss or gain is then a function of the welfare weights attached to the different countries (or, more correctly, to the different individuals who reside in these countries) and of the possibility of income transfers between countries.

[17] Mundell [15] has also examined the effects of tariff preferences on the terms of trade. See also his discussion of the effects of trade taxes [14].

Further, we conclude that it may be rational policy for some countries to push for the establishment of a given customs union, but nevertheless to refrain from joining that union. Increased market share and growth potential outside the union which accrue to a country as a result of the formation of the union will cause that country to elect nonmembership. Finally, we have seen that, viewed as one of several types of tariff policies, customs union may be superior to nonpreferential tariff policy. Not only will there result welfare changes due to variations in the terms of trade, but the tariff revenues collected by a nonprohibitive union may be distributed in a variety of ways which again affect welfare of, and hence the desirability of, the union to a given country.

An interesting by-product of the foregoing analysis is the suggestion that optimum tariff strategy may dictate that some countries remaining outside the union reduce their prevailing tariffs. Much depends upon the extent of tariff warfare prior to formation of the union, and thus upon the dynamics of tariff competition about which very little is known. It is nevertheless intriguing to speculate about the extent to which the existence of the European Economic Community facilitated the Kennedy Round.

References

1. S. W. Arndt, "On Discriminatory vs. Non-preferential Tariff Policies," *Econ. Jour.*, Dec. 1968, 78.
2. G. S. Becker, "A Note on Multi-Country Trade," *Am. Econ. Rev.*, Sept. 1952, *42*, 558–68.
3. J. Bhagwati and H. G. Johnson, "A Generalized Theory of the Effects of Tariffs on the Terms of Trade," *Oxford Econ. Papers*, Oct. 1961, *13*, 225–53.
4. C. A. Cooper and B. F. Massell, "A New Look at Customs Union Theory," *Econ. Jour.*, Dec. 1965, *75*, 742–47.

5. G. A. ELLIOT, "The Theory of International Values," *Jour. Pol. Econ.*, Feb. 1950, *58*, 16–29.
6. F. D. GRAHAM, "The Theory of International Values Re-examined," *Quart. Jour. Econ.*, Nov. 1923, *38*, 54–86.
7. H. G. JOHNSON, "Income Distribution, the Offer Curve, and the Effects of Tariffs," *Manchester School*, Sept. 1960, *28*, 215–42.
8. ———, *International Trade and Economic Growth.* Cambridge 1958.
9. M. C. KEMP, "The Relation Between Changes in International Demand and the Terms of Trade," *Econometrica*, Jan. 1956, *24*, 41–46.
10. R. G. LIPSEY, "The Theory of Customs Union: A General Survey," *Econ. Jour.*, Sept. 1960, *70*, 496–513.
11. J. E. MEADE, *A Geometry of International Trade.* London 1952.
12. L. A. METZLER, "Tariffs, the Terms of Trade, and the Distribution of National Income," *Jour. Pol. Econ.*, Feb. 1949, *57*, 1–29.
13. M. MICHAELY, "On Customs Union and the Gains from Trade," *Econ. Jour.*, Sept. 1965, *75*, 577–83.
14. R. A. MUNDELL, "The Pure Theory of International Trade," *Am. Econ. Rev.*, March 1960, *50*, 67–110.
15. ———, "Tariff Preferences and the Terms of Trade," *Manchester School*, Jan. 1964, *32*, 1–13.
16. J. VANEK, *General Equilibrium of International Discrimination.* Cambridge 1965.
17. ———, *International Trade: Theory and Economic Policy.* Homewood 1962.
18. J. VINER, *The Customs Union Issue.* New York 1950.

Public Utility Pricing and Output Under Risk

By Gardner Brown, Jr. and M. Bruce Johnson*

The theory of publicly produced private goods is still in its infancy—a stepchild of the general equilibrium competitive model of privately produced private goods. Contributors [1], [2], [7], [8] to the theory have generally concluded that production should be pushed to the point where price equals marginal cost. Consequently, any public authority that produces and sells a private good will discharge its responsibility in an optimal manner when it acts as if it were a competitive industry and prices at marginal cost.

The purpose of this paper is to introduce elements of risk into the analysis in ways that seem analytically and empirically plausible. The authority is assumed to be confronted by a product demand function subject to a stochastic disturbance term. The social welfare function that the authority maximizes is the algebraic difference between expected willingness to pay and expected costs. We conclude that the optimal price will always be lower and, with linear demand, the optimal output will generally be higher than their counterparts in the riskless model of traditional theory, regardless of the manner in which the disturbance term enters the analysis. Thus, although the model is a natural extension of the existing peak-load models to a world with risk, the implications for price and output policy are significantly different: the authority will fail to recover its capital or capacity costs unless the environment also includes perfect risk markets. In that event, the revenue lost because of the socially optimal price and output policy of the authority can be recovered through its operations in the futures market for the good or the service.

I. *Riskless Demand*

The demand function of received theory is exact, and the social welfare function is simply the sum of consumers' surplus and total revenue minus the cost of the resources used in the production of the output.[1] Suppose the demand function is written as $Q=X(P)$ where the function is single valued, monotone decreasing, and differentiable. Consumers' surplus plus total revenue is more conveniently expressed as a function of Q; hence $X^{-1}(Q)$ is used to define the welfare index:

$$(1)\qquad W = \int_0^{Q'} X^{-1}(Q)dQ - C(Q').$$

Suppose that total cost is a separable function of operating and capacity cost. Define the two costs per unit of output as b and β respectively; then $C(Q)=(b+\beta)Q$, the type of cost function that has been used extensively in the literature on peak-load pricing problems.[2] Substituting this cost function into the welfare function and maximizing with respect to Q' yields:

$$(2)\qquad P = b + \beta.$$

A geometric exposition of this result is given in Figure 1. The equilibrium levels of the welfare-maximizing price and quantity are determined by the intersection of the demand function and the

* The authors are, respectively, assistant professor of economics, University of Washington and associate professor of economics, University of California, Santa Barbara.

[1] For an excellent discussion regarding the conceptual existence and legitimacy of consumers' surplus, see Patinkin [6].

[2] The average capacity cost per unit of output assumes, of course, that the plant is fully utilized; Hirshleifer [8, p. 814].

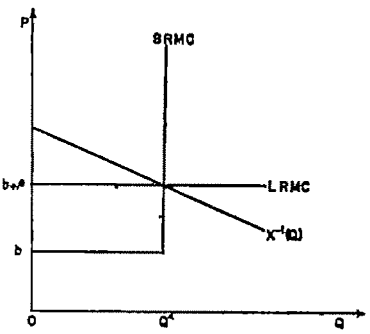


FIGURE 1

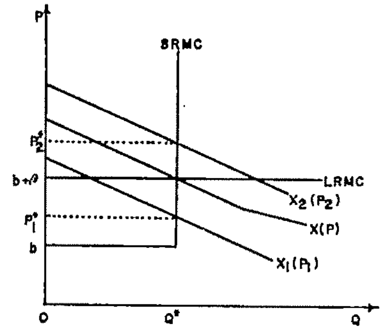


FIGURE 2

long-run marginal cost (LRMC) function. The short-run marginal cost (SRMC) is given by the constant operating cost (b) up to the equilibrium output (Q^*); SRMC becomes vertical at fixed capacity.

Recent contributions to the literature enable us to proceed one step further; for example, see O. E. Williamson [8, pp. 816–27]. If the demand function shifts through a known cycle within a demand planning period, different prices will be charged for peak and off-peak times. Suppose, for the sake of expositional simplicity, $X_1(P_1)$ and $X_2(P_2)$ represent the off-peak and peak demand and each occurs for some fraction of the overall planning and production period (Figure 2). Let these fractions sum to one and be represented by w_1 and w_2.

The new welfare function for the period is then given by

$$(3) \quad W = w_1 \int_0^{Q_1'} X^{-1}(Q_1)dQ_1 + w_2 \int_0^{Q_2'} X^{-1}(Q_2)dQ_2 - w_1 b Q_1' - w_2 b Q_2' - \beta Q_2'$$

where the benefits and costs of each of the two demand functions are weighted by the fraction of the cycle in which they occur. If the capacity of the plant is entirely used in both periods, the equilibrium condition is given by equation (4) and the situation is depicted in Figure 2.

$$(4) \quad \frac{dW}{dQ'} = P_1^* w_1 + P_2^* w_2 - (b + \beta) = 0 \quad \text{for } Q = Q_1' = Q_2'.$$

Equilibrium prices P_1^* and P_2^* clear the market in each of the subperiods (of equal length in this case). The demand-for-capacity curve introduced by Williamson and shown in Figure 2 with the label $X(P)$ determines the equilibrium capacity Q^*.

On the other hand, if the peak and off-peak demand functions are widely separated in the price and output space, the off-peak demand function $X_1(P_1)$ may intersect the SRMC curve on its horizontal rather than its vertical portion. This would imply that the plant is not used to full capacity during off-peak phases of the cycle and would warrant the partial differentiation of equation (3) with respect to Q_1' and Q_2'. The equilibrium conditions then are:

$$(5) \quad P_1 = b \quad \text{and} \quad P_2 = b + \frac{\beta}{w_2}.$$

The off-peak price P_1 is set equal to marginal operating costs, and no capacity charges are levied against this type of use of the facility because capacity is not scarce during the off-peak period. The peak charges, however, include short-run operating charges and full-capacity charges. Although no diagram is presented to illustrate this case, it may easily be visualized.

The preceding peak-load model can be generalized to any number of subperiods

or subcycles with relative ease, and the model seems to be perfectly general since it does not require any particular set of properties for the demand function beyond the usual ones.

II. *Stochastic Demand*

How are the demand functions identified empirically during the peak, interim, and off-peak portions of the cycle? Undoubtedly some plants experience predictable cycles of demand but others, it may be supposed, experience a great deal of random variation juxtaposed on the cycles of peak and off-peak loads. The problem raised by stochastic variation in the demand for a public service is discussed here and can be easily extended to include cases in which random shocks exist in the peak-load and off-peak-load cycle; the conclusions are qualitatively comparable to the simpler situation.

Figure 2 visually represents the problem raised by uncertainty. Suppose that each of the two demand curves depicted occurs 50 per cent of the time during the planning and production period. But further suppose that the plant manager does not know which curve will appear in the first half of the period and which in the second. He must, by assumption, announce a price and output for each half of the period prior to the opening of the period. If he is very fortunate, he will announce a price of P_1^* and will experience the off-peak demand function $X_1(P_1^*)$. However, if the peak demand $X_2(P_1^*)$ occurs, an enormous excess demand will have to be eliminated by nonprice rationing methods of one sort or another. Conversely, had he chosen the price P_2^*, built capacity Q^* in anticipation of $X_2(P_2^*)$, and then the off-peak demand occurred, there would be no sales. Adoption of a different pricing rule, say $P=b+\beta$, would produce only a marginal improvement. Off-peak demand would be met and excess demand at peak times would be reduced, although not to zero. In any event, it is rather unfair to ask such questions of a peak-load model defined for conditions of perfect certainty. Instead, a new welfare-maximizing model embracing conditions of uncertainty will be specified.

In such a model some product or service is provided by a public authority who has been directed to maximize expected social welfare. The authority must assign values to price, P, and capacity output, Z, before the production period begins and before actual demand is known. The problem will be simplified by the assumptions that storage costs are known to be greater than any future price and that disposal costs of unsold output are insignificantly small. The properties of the riskless demand function are:

$$(6)\quad Q = X(P), \quad X'(P) < 0, \quad X^{-1}(0) \text{ finite.}$$

Uncertainty (or, more properly, risk) may enter the demand function in a number of ways. Generally, the demand function may be expressed as

$$(7)\qquad D = D(P, u)$$

where actual demand D is a function of the announced price P and u where u is a continuous random variable. Of particular interest are the additive and multiplicative formulations:

$$(8)\quad D_1 = X(P) + u \quad \text{and} \quad D_2 = X(P)v,$$

which will be analyzed separately.

A. *The Additive Model*

Assume that the plant manager knows both the demand function $X(P)$, and the exact form of the continuous probability density function $f(u)$. The following properties are attributed to $f(u)$:

$$(9)\quad \int_{-\infty}^{+\infty} uf(u)du = 0, \; F(\alpha) = \int_{-\infty}^{\alpha} f(u)du,$$
$$\int_{-\infty}^{+\infty} u^2 f(u)du < \infty.$$

Having announced a price there will be a corresponding $X(P)$ to which nature will

add a random term. Therefore, actual sales of the product or service are

$$(10) \quad S = \begin{Bmatrix} X(P) + u, & u \leq Z - X(P) \\ Z \quad , & u > Z - X(P) \end{Bmatrix}.$$

If the actual demand, $D(P, u)$, is less than capacity output, Z, then D will in fact be actual sales, S. However, if at the announced price, P, a value of u should occur such that $D(P, u)$ is greater than Z, actual sales will be equal to capacity output. Sales cannot be greater than Z although demand may be greater.[3]

Linear cost functions introduced earlier remain in force. The symbol b represents the constant marginal variable cost of producing the good or service in the production period. For many facilities, say the traditional "bridge" in public finance, b may be zero. For other commodities or services it may be the most important cost element of the total cost function. In the case of a public park, b can be interpreted as the variable cost incurred when a campsite is used (cost of protecting its occupants during their stay and the cost of cleaning up after they have departed). The cost for one unit of Z during the demand period is β, and corresponds to a one-period rental rate. Total capacity cost equals βZ, but total variable cost is the product of b and actual sales S during the period. Thus, production (or capacity) cost is known with perfect certainty; operating cost is known subject to a probability distribution.

Following closely the pattern established by earlier riskless theory, the arguments of the welfare function to be maximized include the expected value of willingness to pay, the expected value of variable cost, and the capacity cost known with certainty, once that capacity has been chosen,

$$\begin{aligned}(11) \quad W = {} & E[\text{willingness to pay}] \\ & - E[\text{average variable cost} \cdot \text{sales}] \\ & - E[\text{capacity cost}].\end{aligned}$$

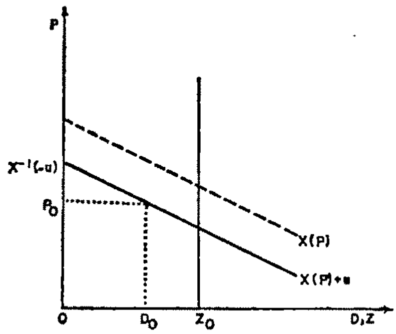


FIGURE 3

The actual welfare function can be constructed piece by piece with the riskless demand function, $X(P)$, serving as a benchmark (Figure 3). Suppose the plant manager arbitrarily chooses values P_0 and Z_0 before he knows the actual value of u. Then if the value of the random variable which turns up is negative and large enough in absolute value, demand will be less than or equal to the capacity of the plant.[4]

The relevant areas under the demand function are consumers' surplus and total revenue. The expectation of the sum of the areas when demand including the shock term is less than capacity Z_0, is (12).

Note that the expectation taken from $-\infty$ to $+\infty$ is not strictly correct be-

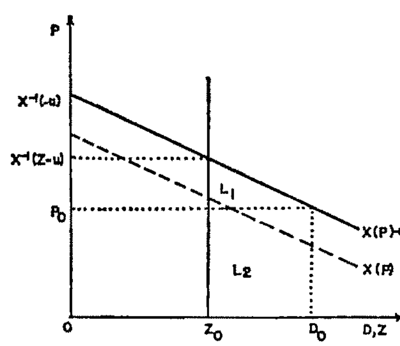


FIGURE 4

[3] In other words, capacity output is chosen before u is known. Actual output is chosen after u is observed by setting actual output equal to the smaller of capacity output and actual demand.

[4] The intercept of the demand function $X(P)+u$ is computed by setting demand equal to zero and solving for the inverse, $X^{-1}(-u)$.

cause it includes those occasions when demand will be greater than capacity output. The overestimate is accounted for in the loss functions below.

When u takes on large positive values, actual demand will be greater than output and losses in consumers' surplus and in total revenue occur. This is illustrated in Figure 4 where the explicit assumption is made that when output falls short of demand, available output is distributed to individual customers on the basis of their valuation of the marginal unit of the good.[5] The losses in surplus and revenue are denoted L_1 and L_2 and their expectations are given by (13).[6]

Since $E(L_1)$ and $E(L_2)$ are positive only when $u > Z - X(P)$, this furnishes the lower limit of integration on the integral involving u.

Expected sales (14) equal the expectation of demand, $X(P)+u$, over the entire range of the random variable minus the truncated mean whenever demand exceeds capacity.[7]

The product $bE[S]$, together with βZ, gives the cost arguments which contribute negatively to welfare. After some rearrangements of terms, the explicit expression for W is obtained by subtracting expected losses and costs from expected gross willingness to pay in (15).

When W is differentiated with respect to price and capacity, the first-order conditions for maximization of the welfare function are in (16).

It is easily shown that the second-order conditions for maximization are also satisfied. After simplification the first-order conditions indicate that price equals short-run marginal operating cost and optimal capacity should be chosen such that marginal capacity cost is equal to the truncated mean of the difference between the willingness to pay and the actual price for the marginal disappointed purchaser of the commodity.[8]

These results differ sufficiently from the equilibrium conditions of the riskless model to warrant further discussion. The introduction of risk into the analysis requires the authority to choose levels of price and output simultaneously. Optimal price is below LRMC and, more precisely, is driven down to SRMC because of the nature of the maximand. Given that the optimal price is below the riskless optimal price, the law of demand would seem to suggest that optimal output will be greater than equilibrium output in the riskless model.

The necessary conditions given in (15) and (16) cannot readily be solved for the optimal output since the integral in (16) uniquely depends on the probability density function $f(u)$ and the demand function $X(P)$. Consequently, for illustrative purposes a linear demand function is introduced and it will be shown that optimal output is unambiguously greater in this model with risk than it is in the traditional deterministic model.

The strategy will be to set Z equal to riskless Z in (16) and show that the expression is greater than rather than equal to zero. Clearly, the implication is that the riskless capacity level is too small, that the marginal contribution of capacity to the welfare function is still positive, and that for (16) to be satisfied, optimal capacity must be greater than riskless capacity for all admissible demand functions.

Let $X(P)=A+BP$ where $A>0$ and $B<0$. From the riskless model we know

[5] We assume that purchasers are ranked according to their willingness to pay for the product even though all pay the same price P_0.

[6] The value of price at the intersection of the vertical capacity curve and the demand function $X(P)+u$ is obtained by setting $X(P)+u=Z$ and solving for $P=X^{-1}(Z-u)$.

[7] For a lucid discussion of expected demand in a monopoly model see E. Mills [5, pp. 82–99], an analysis to which the present authors owe much.

[8] We have assumed that those purchasers with the greatest willingness to pay are serviced first. If any other assumption is made about the actual distribution of the commodity or service to the potential buyers, the shadow price of an extra unit of output obtained through additional capacity would be higher and would yield an even larger value of Z.

$$(12)\quad \int_{-\infty}^{\infty} f(u) \int_{P}^{X^{-1}(-u)} [X(P) + u]dPdu + PX(P) \qquad \text{for } X(P) + u \leq Z$$

$$(13)\quad \left.\begin{aligned} E[L_1] &= \int_{Z-X(P)}^{+\infty} f(u) \int_{P}^{X^{-1}(Z-u)} [X(P) + u - Z]dPdu \\ E[L_2] &= \int_{Z-X(P)}^{+\infty} f(u)P[X(P) + u - Z]du \end{aligned}\right\} \text{for } X(P) + u > Z.$$

$$(14)\quad E[S] = \int_{-\infty}^{+\infty} f(u)[X(P) + u]du - \int_{Z-X(P)}^{+\infty} f(u)[u - (Z - X(P))]du.$$

$$(15)\quad \begin{aligned} W = &\int_{-\infty}^{\infty} f(u) \int_{P}^{X^{-1}(-u)} [X(P) + u]dPdu + PX(P) \\ &- \int_{Z-X(P)}^{\infty} f(u) \left\{ \int_{P}^{X^{-1}(Z-u)} [X(P) + u - Z]dP + P[X(P) + u - Z] \right\} du \\ &- b\left\{ X(P) - \int_{Z-X(P)}^{\infty} uf(u)du + [Z - X(P)] \int_{Z-X(P)}^{\infty} f(u)du \right\} - \beta Z \end{aligned}$$

$$(16)\quad \begin{aligned} \frac{\partial W}{\partial P} &= PX'(P)F[Z - X(P)] - bX'(P)F[Z - X(P)] = 0, \\ \frac{\partial W}{\partial Z} &= \int_{Z-X(P)}^{\infty} f(u)[X^{-1}(Z - u) - b]du - \beta = 0. \end{aligned}$$

---

that output will be determined by the intersection of the riskless demand function and the LRMC function. Consequently, $P=b+\beta$ and $\bar{Z}=X(b+\beta)$ where the bar denotes riskless Z. Substituting for $X(P)$ and Z in (16) gives:

$$(17)\quad \beta \int_{B\beta}^{\infty} f(u)du - \frac{1}{B} \int_{B\beta}^{\infty} uf(u)du = \delta.$$

It must be proved that

$$\delta > 0 \qquad \text{for } [-A/(b + \beta)] < B < 0,$$

the admissible set of values for B. The first inequality insures that demand is positive and that a solution exists in the riskless case with which the comparison is being made; the second obviously restricts the investigation to downward-sloping demand functions. By inspection it is clear that the limit of δ as B goes to zero from below will be positive; similarly, δ goes to zero from above as $B \to [-A/(b+\beta)]$. Consequently, optimal capacity must be greater than riskless capacity: $Z^* > \bar{Z}$.

B. *Multiplicative Model*

Assume that actual demand is defined as $D=X(P)v$ where v is a positive random variable distributed according to the continuous probability density function $g(v)$ where $E[v]=1$. Figures 5 and 6 indicate that the intercept $X^{-1}(0)$ is independent of the value taken by the disturbance term.

The welfare function comprising expected consumers' surplus and expected costs is constructed along lines parallel to the additive case discussed above. From Figure 5 the expected gross willingness to

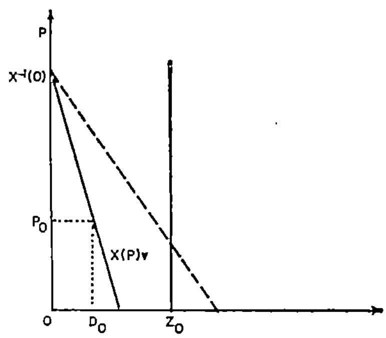


FIGURE 5

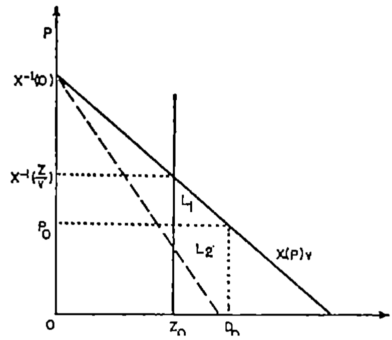


FIGURE 6

pay, given P_0 and Z_0, is

$$(18) \quad \int_0^\infty g(v) \int_P^{X^{-1}(0)} vX(P)dPdv + PX(P).$$

However, since v may take on values that lead to excess demand for arbitrarily given P_0 and Z_0, the expected losses in consumers' surplus and total revenue labeled L_1 and L_2 in Figure 6 must be computed.[9]

$$(19) \quad E[L_1+L_2] = \int_{Z/X(P)}^\infty g(v) \int_P^{X^{-1}(Z/v)} [X(P)v-Z]dPdv + \int_{Z/X(P)}^\infty g(v)P[X(P)v-Z]dv.$$

Expected sales in this model equal the expectation of demand over the entire domain of the random variable, minus the truncated expectation of excess demand whenever demand exceeds capacity (i.e., whenever $v > Z/X(P)$).

$$(20) \quad E(S) = X(P) - \int_{Z/X(P)}^\infty g(v)[X(P)v - Z]dv.$$

Now form the welfare function by subtracting the sum of expected losses and costs $E[L_1+L_2]+bE[S]+\beta Z$ from the expected gross willingness to pay in equation (18). Necessary conditions for maximization of the welfare function are that the partial derivatives with respect to P and Z vanish. This implies, after simplification,

$$(21) \quad p = b \quad \text{and}$$

$$\int_{Z/X(P)}^\infty g(v)\left[X^{-1}\left(\frac{Z}{v}\right) - b\right]dv = \beta.$$

Notice that this result is qualitatively the same as that obtained in the additive case. Optimal price equals marginal operating cost and optimal capacity output should be chosen such that the truncated expectation of the willingness-to-pay of the marginal disappointed user is equal to marginal capacity cost. Just as in the additive case, optimal price is lower under risk than it is in the traditional deterministic model[10] where $P = b + \beta$.

Given the assumption of a linear demand function, we wish to show that if optimal capacity output under risk were equal to riskless capacity, the partial derivative of

[9] The value of price at the intersection of the vertical capacity curve and the demand function is obtained by setting $X(P)v = Z$ and then solving for $P = X^{-1}(Z/v)$.

[10] The invariance of these results with respect to the additive and multiplicative formulations of the model is not without interest in view of established results from profit maximizing under uncertainty. "The optimal price under multiplicative uncertainty is always *greater*, and under additive uncertainty always *less*, than the optimal price set in the equivalent deterministic monopoly models" (emphasis added); Karlin and Carr [4, p. 160].

the welfare function with respect to capacity Z (i.e., equation 21) will be greater than zero thus signifying that capacity should be expanded until the contribution to welfare is zero at the margin. Let $X(P)=A+BP$ as before. Then,

$$(22)\quad \left(\frac{A}{B}+b+\beta\right)\int_{v_0}^{\infty}\frac{g(v)}{v}\,dv - \left(\frac{A}{B}+b\right)\int_{v_0}^{\infty}g(v)dv - \beta = \delta,$$

$$\text{where } v_0 = \frac{A+B(b+\beta)}{A+Bb}$$

under the assumption that $Z=\bar{Z}=X(b+\beta)$. The admissible values of B are

$$[-A/(b+\beta)] < B < 0.$$

As $B\to 0$, $v_0\to 1$, and $(A/B)\to -\infty$. But (22) can be rewritten:

$$(23)\quad \delta = \left(\frac{A}{B}+b\right)\int_{v_0}^{\infty}g(v)\left(\frac{1-v}{v}\right)dv - \beta\left[1-\int_{v_0}^{\infty}g(v)\,\frac{1}{v}\,dv\right]$$

Since the truncated expectation of $[(1-v)/v]$ will invariably be negative, the limit of δ as B goes to zero from below will be positive. As $B\to[-A/(b+\beta)]$, $-B\beta\to(A+Bb)$, $v_0\to 0$ and $\delta\to 0$ from below. Thus it is obvious that $Z^*>\bar{Z}$ unless demand is very elastic and β very large. Consequently, optimal output under conditions of risk will generally exceed the corresponding output from the traditional riskless model given linear demand. This is qualitatively similar to the result derived in the additive case.

C. *A Special Case: The Rectangular Distribution in the Additive Model*

For illustrative purposes consider the example of the linear demand function $X(P)=A+BP$ and the rectangular distribution[11] where $f(u)=1/2\lambda$. This specification enables us to integrate the first-order condition and to solve the resulting quadratic for the optimal value of output.

$$(24)\quad Z^* = A + Bb + \lambda \pm 2(-\lambda B\beta)^{1/2}$$

The negative root is clearly the economically relevant one here, inasmuch as the positive root would indicate an output greater than the maximum admissible value of actual demand.

It is instructive to compare the value of Z^* derived here with the value of output $\bar{Z}$ derived from the comparable riskless case. In the latter, $P=b+\beta$ implies $\bar{Z}=A+B(b+\beta)$. The comparison is achieved by examining the difference

$$(25)\quad Z^* - \bar{Z} = \lambda - 2(-\lambda B\beta)^{1/2} - B\beta$$

Multiply both sides by the positive constant λ. Then

$$(26)\quad \lambda(Z^* - \bar{Z}) = [\lambda - (-\lambda B\beta)^{1/2}]^2.$$

Since $\lambda>0$, it follows that $Z^*-\bar{Z}\geq 0$. If $\lambda=-B\beta$, $Z^*=\bar{Z}$. In general, output under conditions of risk is larger than under perfect certainty; given a linear demand function subject to an additive disturbance term distributed according to a rectangular probability law, optimal output exceeds expected sales.[12]

The effects of changes in the parameters in the solution for Z^* can be shown by differentiating (24) partially with respect to the parameters. Z^* will increase with increases in A and B and will decrease with increases in b and β. Nor is there any difficulty interpreting these partials: in geometrical terms, as the demand curve becomes steeper in the traditional price-quantity diagram, optimal output in-

[11] To insure nonnegative demand we assume $\lambda\leq A+Bb$. Thus, at the optimal price $P=b$ actual demand $X(P)+u$ is greater than $(u>-\lambda)$ or equal to $(u=-\lambda)$ zero.

[12] In this special case expected sales equal riskless sales. To show this, substitute Z^* into the expression for expected sales.

creases. As marginal operating and capacity costs increase, optimal output will be reduced. Finally, the first partial derivative of Z^* with respect to λ is uniformly positive for $-B\beta<\lambda<A+Bb$. As $\lambda\rightarrow -B\beta$, $Z^*\rightarrow \bar{Z}=A+Bb+B\beta$, riskless demand. In general, optimal output will be greater than riskless output for admissible values of the parameter λ.

III. *The Way Back*

The formulation of the welfare function in a riskless neoclassical world insures that net revenue is identically equal to zero. By contrast, the formulation of a comparable social welfare function when the demand function is stochastic provides an optimal price and output which insures that the public agency will have negative net revenues from its activities; short-run operating costs will be recovered but capacity cost will not. There are precedents in the literature for these conclusions; for example, see the work of Professor H. Hotelling in the area of decreasing-cost industries [3]. However, our extension of these results by a conceptually distinct analysis to usually well-behaved cost curves may be unique. We suspect, although from admittedly casual empiricism, that the model we have proposed may be surprisingly consistent with observed behavior in the so-called real world. The practice of subsidizing capacity costs from general funds and covering operating costs out of user charges is not an unusual practice, in other words.

The analysis presented here indicates that optimal policy calls for a lower price and generally a larger quantity than would be obtained from the traditional model in a world without risk. Suppose perfect risk or futures markets exist in the world of our model and that the public authority chooses price and quantity according to the dictates of the risk model presented above. Potential speculators would certainly recognize that in particular periods the stochastic disturbance term could produce a demand so low that excess capacity would exist at the announced price and, consequently no benefit would accrue to those individuals who held the right or reservation to buy the service or good during that period. The right would be worthless.

By the same argument, the disturbance term could also produce demand so large in other periods that the speculator who held the right to buy the good or service at the public authority price, could sell his right in the spot market for the premium $X^{-1}(Z-u)-b$. If speculators calculate the expected value of this premium they will conclude, of course, that it is equal to β since that was the way the authority initially determined output.

If the public authority wants to cover its capacity costs, the prescription is plain. Enter the futures market and sell Z^* rights to the good or service at a price of β each. Speculators will be willing to pay that price, but no more. Thus, on the average, the price per unit of output paid by the consumer will be equal to $b+\beta$, the actual price that holds in the riskless model even though optimal output is greater than riskless output![13]

References

1. M. Boiteux, "La tarification des demandes en point: application de la théorie de la vente au coût marginal," *Revue Générale de l'Électricité*, Aug. 1949, *58*, 321–40; translated as "Peak-Load Pricing," *Jour. Bus.*, April 1960, *33*, 157–79; reprinted in James R. Nelson, ed., *Mar-*

[13] Notice that if the authority had produced capacity equal to the smaller riskless output, demand would exceed capacity a greater percentage of the time, the expected premium would be larger, the amount speculators would offer for rights would exceed marginal capacity cost β, effective price would exceed long-run marginal cost, and the public authority would report a positive profit in violation of the rules of the competitive game.

ginal Cost Pricing in Practice, Englewood Cliff 1964, pp. 59–90.
2. J. Hirshleifer, "Peak Loads and Efficient Pricing: Comment," *Quart. Jour. Econ.*, Aug. 1958, *72*, 451–62.
3. H. Hotelling, "The General Welfare in Relation to Problems of Taxation and of Railway and Utility Rates," *Econometrica*, July 1938, *6*, 242–69.
4. S. Karlin and C. R. Carr, "Prices and Optimal Inventory Policy," in K. J. Arrow, S. Karlin and H. Scarf, ed., *Studies in Applied Probability and Management Science*, Stanford, California, 1962, 159–72.
5. E. Mills, *Price, Output, and Inventory Policy*, New York 1962.
6. D. Patinkin, "Demand Curves and Consumer's Surplus," in *Measurement in Economics; Studies in Mathematical Economics and Econometrics in Memory of Yehuda Grunfeld by Carl F. Christ and others*. Stanford, 1963, 83–112.
7. P. O. Steiner, "Peak Loads and Efficient Pricing," *Quart. Jour. Econ.*, Nov. 1957, *71*, 585–610.
8. O. E. Williamson, "Peak-Load Pricing and Optimal Capacity Under Indivisibility Constraints," *Am. Econ. Rev.*, Sept. 1966, *56*, 810–27.

The Structure of the Money-Expenditures Relationship

By SAM PELTZMAN*

Some recent studies (e.g., [1] [6]) have found it useful to treat the demand for money within a stock-adjustment model. The intended change in money balances held is represented as an attempt to adjust the stock of money held currently to the desired or long-run equilibrium stock demanded. One form that these adjustments take will be expenditures on goods and services. In fact, its relationship to total expenditures or income is the most important basis of the economist's interest in the demand for money. Therefore, a model which explains the demand for money successfully should be useful in explaining the behavior of expenditures on goods and services. This paper investigates the empirical usefulness of a stock-adjustment model in explaining the relationship of money to expenditures.

I. *Stock-Adjustment in the Money-Expenditures Relationship*

Traditional monetary theory has used a stock-adjustment model of a restricted form to describe the money-expenditures relationship; we shall simply drop some of these restrictions. As a starting point, consider the simple quantity theory of money. For our purposes, we can write the behavioral relationship it generates in the linear form:

$$\text{(1)} \quad \left(d\left(\frac{E}{M}\right) \Big/ dt\right)^* = a[(E/M)^* - (E/M)], \quad \text{where}$$

E = total expenditures

M = the quantity of money

a = a constant coefficient of adjustment

* denotes expected or intended value.

This says that intended changes in the ratio of expenditures to money—i.e., velocity—occur in response to a discrepancy between desired and actual velocity. The simple quantity theory specifies that desired velocity is a constant, assumes equality of intended and actual changes in velocity, and sets the adjustment coefficient equal to one. That is, it assumes that there is continuously complete adjustment to any discrepancy between desired and actual velocity. While we shall relax the constant-desired-velocity assumption, our main concern will be to relax the assumption of continuously complete adjustment of actual to desired velocity.

Economic theory does not, in fact, suggest that such adjustment will be continuously complete. Some costs of adjustment may be sufficiently reduced by an investment of time to make the investment worthwhile. For example, consider an individual who has received an unanticipated increase in his money balances in a world with information costs. He will react to such costs in goods markets by spending *some* time in sampling them, and thus he will spend *some* time in expending the unanticipated money balances. Of course,

* The author is assistant professor of economics, University of California, Los Angeles. He is indebted to Dennis Jesmok and Michael Carney for research assistance, and to the Bureau of Business and Economic Research, University of California, Los Angeles for financial assistance.

given the time units over which economic data are measured, the assumption that adjustment is continuously complete may be useful empirically. However, the most direct way of determining the empirical usefulness of this assumption is to estimate the adjustment coefficient in (1) directly;[1] this is what I shall do.

Some of the peculiar empirical consequences of assuming continuously complete adjustment of velocity provide further motivation for questioning the assumption. In practice, the assumption leads to tests of the quantity theory which regress levels or changes in some type of expenditure on levels or changes in money. This is the approach of Friedman and Meiselman in [4]. Here is a typical result of that approach for quarterly data from 1953 I to 1967 II:

$$\text{(E1)} \quad \Delta E_t = .852\times 10^{-2} + .406\Delta M_{t-2}$$
$$(.166\times 10^{-2}) \quad (.116)$$
$$r = .423, \quad \text{s.e.} = .680\times 10^{-2}, \quad n = 58,$$

where standard errors of the coefficients are in parentheses, E is consumption, M is currency plus demand and time deposits at commercial banks, and the changes are of natural logarithms. The relationship between changes in money and changes in expenditures is significant, but the parameters of the relationship include a significant time-trend in consumption and a money elasticity of expenditures significantly less than the unity implied by the simple quantity theory. The failure of (E1) and its variants[2] to show equiproportionality between money and expenditure changes does not, however, vitiate the significant overall relationship between the two. It suggests, though, that we may improve on the structure of (E1).

I proceed, then, by retaining the essential quantity theory framework contained in (1), but I shall not constrain the adjustment coefficient to unity. First, treat all variables in (1) as natural logarithms,[3] so that it may be rewritten:

$$(2) \quad (dE/dt)^* = a[(E/M)^* - (E/M)] + (dM/dt)^*.$$

This breaks up an intended expenditure change into an expected change in money plus an intended change in velocity. At this point I simplify by assuming that $(dM/dt)^*$ is a constant, so all deviations of $(dE/dt)^*$ from its long-run mean are regarded as adjustments to a discrepancy between desired and actual velocity. We then specify the money-demand function, in terms of desired velocity, in the log-linear form:

$$(3) \quad (E/M)^* = b + c(E/P) + ei \quad \text{where}$$
$$(E/P) = \text{real expenditures}$$
$$i = \text{the rate of interest.}$$

The expected values of c and e are matters of some controversy. The sign of e can be expected to be positive, and most studies of the demand for money find that to be the case. However, the estimated size of the interest rate response of velocity has varied considerably. (See, e.g., [2] [7].) Controversy about c goes even further—

[1] Some empirical evidence on the demand for money already casts doubt on its usefulness. Hamburger [6, p. 621] concludes that "... households seem to be somewhat slow in adjusting their money balances to changes in asset yields and the level of prices. These adjustments are generally not completed within the quarter in which the disturbance occurs, ..."

[2] No improvement over (E1) is obtained if the relevant money change is viewed as a weighted average of past changes. When we added other lagged money changes, they were never significant, and the sum of their coefficients was always close to the .4 for the single lagged value in (E1). Friedman and Meiselman's results [4, p. 239] are similar. They use absolute changes, and the sum of coefficients never approaches the actual consumption velocity of money.

[3] From this point on, all variables in upper case are natural logarithms, those in lower case are absolute values. When a ratio is in parentheses, it may be treated as a single variable. Hence, e.g., $(E/M) = E - M$. Also note that the variable on the left-hand side of (2) is the intended change in expenditures, not the change in intended expenditures. These two will be equal only if $a = 1$.

there is no conventional wisdom even about its sign, at least for the post-World War II period from which we shall draw our data. Up to this period, the secular trend of velocity had been downward, indicating that money is a superior good (c is negative). However, the postwar period has seen a substantial rise in velocity, and this has led to conjecture that the income elasticity of money has fallen below unity. The reason most frequently adduced for this changed elasticity is the growth of money substitutes in the postwar period. Since interest rates have risen in this period, the postwar rise in velocity may be accounted for by a sufficiently large value of e. However, the postwar rise in velocity may be explained by neither income nor interest rates. For example, Friedman and Schwartz [5, Ch. 12] have argued that most of the postwar rise in velocity may be attributed to the gradual recognition by money holders that the economic environment has become more stable. Such increased stability lowers expected transaction costs for nonmoney wealth and so lowers the demand for money relative to other forms of wealth. While we shall not test Friedman and Schwartz's conjecture directly, we do want to allow for the more general point implied thereby: factors other than income and interest rates and specific to the postwar period may be responsible for the postwar rise in velocity. The validity of this general class of argument can be tested by including a secular-trend term in the desired velocity equation for the postwar period. Since trend is a proxy for the class of "other" factors, the presence of significant secular trend in postwar velocity would mean that income and interest rate movements cannot explain fully the behavior of postwar velocity.

Both income and interest rates have postwar secular trends of their own. In order to measure the independent influence of trend accurately we would like to reduce the problems associated with the multicollinearity that is introduced by common time trends. Therefore, I proceed as follows: first, add a trend term (ft) to the right-hand side of (3), and substitute this expression for desired velocity into (2) to get

$$(4) \quad (dE/dt)^* = a[b + c(E/P) + ei + ft - (E/M)] + (dM/dt)^*.$$

Regrouping terms, this becomes

$$(4') \quad (dE/dt)^* = [ab + (dM/dt)^*] + ac(E/P) + aei + aft - a(E/M).$$

The bracketed term on the right-hand side of (4′) is a constant, while empirical counterparts of the remaining terms all have secular trends. However, by taking the first derivative of (4′), we get

$$(4'') \quad \frac{d[(dE/dt)^*]}{dt} = 0 + ac\frac{d(E/P)}{dt} + ae(di/dt) + af - a\frac{d(E/M)}{dt}.$$

In this form, effects attributable solely to trend are expressed in a constant, af, which cannot, of course, reflect movements in the other variables. af should equal zero if postwar movements in velocity can be attributed entirely to movements in income or interest rates; otherwise af will be positive.

II. *Empirical Estimates*

To make (4″) operational, it is assumed that intended and actual changes are equal. Aggregated quarterly data for the United States from 1953 I through 1967 II[4] are used to construct discrete approxi-

[4] This period is chosen so as to exclude effects of the Federal Reserve price support program for government bonds. This program might, for example, have lowered the demand for money by lowering the expected variance of bond prices.

mations to the variables in (4″), and (4″) is then estimated by least squares. The results are in Table 1.

These regression estimates employ two expenditure totals—personal consumption and net national product. Personal consumption is used as a proxy for "permanent" income. The inclusion of consumption in the regressions alongside "measured" income is meant to test the hypothesis that permanent, rather than measured, income is the relevant income variable in the demand-for-money equation. This hypothesis has received empirical support in other studies [1] [3] [8]. These two expenditures totals are used in conjunction with three money totals: the conventionally defined money supply (currency plus demand deposits), conventional money plus commercial bank time deposits, and conventional money plus commercial bank time deposits, mutual savings bank and savings and loan association deposits. A recursive model is assumed in which it takes one quarter for money holders to respond to a discrepancy between desired and actual velocity; hence all independent variables are lagged one quarter.

None of the regressions in Table 1 reports an interest rate coefficient. The reason for this appears in the last column. The interest rate variable (yield on long-term government bonds) does not have a significant coefficient in any of the estimates of (4″). Wherever its coefficient is "suggestive," it is of the wrong-sign. On the basis of this evidence, we deleted the interest rate from our model.[5] This poor performance of an interest rate variable tends to support the view that the interest rate variable is an unimportant argument in the demand-for-money function, but we repeat that some other studies fail to support this view. I will discuss the other aspects of the now simplified demand-for-money equation which is implicit in the results after discussing the findings for the speed of the expenditure-adjustment to changes in money.

First, the stock-adjustment model seems to be successful when applied to the money-expenditures relationship: half the variance of the dependent variable on one definition, and a third on another, can be explained by the variables in that model,[6] and the parameter in which we are most interested, the stock-adjustment coefficient (i.e., the coefficient of actual velocity) is is always highly significant and of the expected sign. Moreover, this coefficient is almost always significantly less than unity in absolute value, which indicates that, for quarterly data, the effects of a money change in one period are not completed within any one period of similar length. On the other hand, the effects do not spread over very many periods: 40 to 60 per cent, depending on the expenditure definition, of the discrepancy between desired and actual velocity is made up in one quarter. The total length of effectiveness of a change in money depends on the degree of inflation induced by the money change. It takes longer for adaptation to a money change the greater the difference

[5] In order to save space, we do not report two sets of regressions, one with, the other without, an interest rate variable. However, we point out that none of the coefficients of the other independent variables is changed much if the interest rate variable is included. For example, in the estimate on which the interest rate has the greatest effect, E5, the coefficient (standard error) of the real expenditures and velocity variables are .023 (.189) and −.565 (.214) respectively.

We experimented with a short rate (treasury bill yield) to no avail. The t-ratios for the coefficients of the short rate were 0.28, 0.85, and 1.12 for the three consumption regressions in order, and −1.28, −0.78 and −0.48 for the net national product regressions. The highest (positive) point-estimate for the interest rate elasticity of desired velocity (measured about the mean of interest rate) is .014, this for the short rate in $E4$.

[6] Further, E3 is superior to E1 when E1 is estimated in second differences to render it comparable to E3. The standard error of estimate of E1 in that form is 1.030×10^{-2}.

TABLE 1—REGRESSION RESULTS, 1953 I THROUGH 1967 II (58 OBSERVATIONS)

Estimate Number and Description	Constant (×100)	Real Expenditures	Expenditures-Money Ratio	R	$s.e.$ (×100)	$D.W.$	t for Interest Rate
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Expenditures defined as consumption, and money defined as:							
E2: currency plus demand deposits	.750 .149	−.386 .137	−.479 .157	.675	.721	2.19	−0.60
E3: currency plus demand deposits plus commercial bank time deposits	.454 .148	−.421 .126	−.400 .121	.685	.712	2.15	0.02
E4: currency plus demand deposits plus commercial bank time deposits plus deposits at mutual savings banks and savings and loan associations	.261 .183	−.372 .141	−.446 .148	.673	.723	2.19	0.17
Expenditures defined as Net National Product, and Money defined as:							
E5: Currency plus demand deposits	.494 .166	−.097 .184	−.564 .220	.579	1.00	1.99	−1.95
E6: Currency plus demand deposits plus commercial bank time deposits	.121 .169	−.098 .149	−.562 .161	.625	.958	2.17	−1.37
E7: Currency plus demand deposits plus commercial bank time deposits plus deposits at mutual savings banks and savings and loan associations	−.218 .216	.043 .171	−.694 .184	.640	.944	1.89	−1.76

Note: The dependent variable is the quarterly first difference of the quarterly change in the natural log of expenditures. The independent variables are quarterly first differences of the natural log of the indicated variables. All independent variables are lagged one quarter. Coefficients are shown above their standard errors.

R is the multiple correlation coefficient, $s.e. \times 100$ is the standard error of estimate in percentage points, $D.W.$ is the Durbin-Watson statistic (which indicates no significant serial correlation in the residuals of any regression in the table). The last column gives the ratio of the coefficient of the interest rate variable to its standard error if the regression had been expanded to include it.

Sources of Data: Personal consumption expenditures, net national product, real personal consumption expenditures, real net national product all from *Survey of Current Business*. Money supply data and interest rate (yield of long term government bonds) data from *Federal Reserve Bulletin*. All data are seasonally adjusted with the exception of mutual savings bank and saving and loan deposits; preliminary analysis indicated that neither of these series had any strong seasonal pattern. Flow variables are annual rates.

between the nominal and real expenditure change produced by the money change. This asymmetry arises on account of feedback between desired velocity and real expenditures.[7]

I have illustrated both the effects of this asymmetry and the speed with which changes in money are diffused, in Table 2. The effects of two hypothetical money changes are presented for the limiting cases of pure inflation and no inflation. Column (2) in the table shows the change in expenditures in successive periods after the initial change in money. These numbers are derived from the weights in one of the consumption regressions in Table 1, (E3) except that the trend-factor, measured by the constant term in E3, is ignored. This is done because the trend-factor may be specific to the postwar period, and the examples should be general. The numbers in column (2) show that the length of effectiveness of a change in money and the equilibrium change in nominal expenditures are both increased substantially when there is inflation.[8] The length of effectiveness of a money change is, however, affected little by making the change gradual because the adjustment coefficient is so large.

Finally, we compare the results in Table 1 with the more simple model in E1. There is close agreement between the money-elasticity of expenditures in E1 and the money-elasticity of expenditures *in the quarter of maximum effectiveness* in E3; both are about. 4.[9] The generally good fit of E1 is attributable to the high value of the adjustment coefficient in E3. E3 improves on E1 by showing the total effect of one quarter's change in money and how long it takes to achieve it.

III. *Implications for the Demand for Money and the Post-War Rise in Velocity*

Each regression in Table 1 contains an implicit estimate of the demand-for-money function. These estimates are notable for the decisive way in which they support the general class of argument typified by Friedman and Schwartz's reconciliation of the postwar rise in velocity with its long-run downward trend. Their particular explanation, we recall, is that secularly "... changing views about economic stability can account for the magnitude of the changes in the velocity of money" [5, p. 675]. My results cannot, of course, discriminate between this explanation and others which rely on exogenous trend-related phenomena. However, these results do imply strongly that interest rate movements cannot explain the postwar rise in

[7] To illustrate: an increase in money will initially produce an excess of desired over actual velocity. If there is unemployment, both real and nominal expenditures will rise in response to this discrepancy. This reduces the discrepancy from two sides: (1) actual velocity rises, (2) desired velocity falls; it is negatively related to real expenditures under one (and, it is argued, the relevant) expenditure definition. If there is full employment, however, real expenditures do not rise, and the gap is closed only on account of the rise in actual velocity. Consequently, the remaining discrepancy is larger in successive periods, and it takes longer to eliminate the initial discrepancy. This implication of the model deserves further study, since Table 2 shows that the asymmetry can be empirically important.

[8] The insignificant real-expenditures coefficient in the net national product regressions in Table 1 means that there are no substantial differences due to inflation in the response of that expenditure total to changes in money.

[9] But there is a one quarter difference in timing. Experimentation with alternative lags in Table 1 regressions proved futile; the one quarter lag model always maximized explanatory power. For contemporaneous data, the typical actual velocity coefficient was insignificantly positive, reflecting the effect of common forces operating on the dependent and independent variables. For two and three quarter lags, this coefficient was negative, but insignificant. The magnitudes of this coefficient, interestingly, were almost always close to that predicted by our model. If a one quarter lag model is really appropriate, the E3 adjustment coefficient for any lag would be bracketed by the two entires for Example 1 in column (2) of Table 2 for that lag, divided by 10. For E3 estimated with a two and three quarter lag, the stock adjustment coefficients were .083 and .014 respectively. Similar results obtained for all of the regressions in Table 1, and this tends to support the choice of a one quarter lag model.

TABLE 2—ILLUSTRATIVE EFFECTS OF INCREASES IN MONEY ON CONSUMPTION

Example	Number of Quarters after Initial Change in Money (1)	Percentage Change in Consumption (2)	Values of Dependent Variables in Natural Logs: Real Consumption (3)	Values of Dependent Variables in Natural Logs: Actual Velocity (4)
	Case (a) Less-than-full employment, no inflation			
1. Ten per cent increase in money in one quarter	0	0	0	0
	1	4.00	0	−.1000
	2	0.72	.0400	−.0600
	3	0.13	.0472	−.0528
	4	0.02	.0484	−.0516
	5	Nil	.0486	−.0514
	∞	Nil	.0487	−.0513
2. Ten per cent increase in money spread equally over two quarters	0	0	0	0
	1	2.00	0	−.0500
	2	2.36	.0200	−.0800
	3	0.42	.0436	−.0564
	4	0.08	.0478	−.0522
	5	Nil	.0486	−.0514
	∞	Nil	.0487	−.0513
	Case (b) Full employment, pure inflation			
1. Ten per cent increase in money in one quarter	0	0	0	0
	1	4.00	0	−.1000
	2	2.40	0	−.0600
	3	1.44	0	−.0360
	4	0.86	0	−.0216
	5	0.52	0	−.0130
	∞	Nil	0	0
2. Ten per cent increase in money spread equally over two quarters	0	0	0	0
	1	2.00	0	−.0500
	2	3.20	0	−.0800
	3	1.92	0	−.0480
	4	1.15	0	−.0288
	5	0.69	0	−.0173
	∞	Nil	0	0

Source: see text; all weights are from E3, Table 1. The initial values of real expenditures and actual velocity are both assumed to equal 1. All percentages are continuously compounded quarterly rates. The money and expenditure changes are to be interpreted as deviations from means.

velocity, nor can the phenomenon be explained by some change in the velocity-income relationship.

I concentrate on the consumption regressions, though many of our conclusions about the postwar behavior of velocity do not depend critically on this choice. This choice of regression form is made on the basis of goodness of fit: the consumption-proxy for permanent income is the better expenditures measure (the standard error of estimate is the relevant basis of comparison). Further, the ordering of the coefficients of real consumption and real

net national product is consistent with a permanent income explanation of long-run velocity movements: the latter is smaller absolutely than the former because of the variance of the "transitory" components in measured income. (See [3].) These results hold for any of the three money totals employed, and indeed there is little choice among the three.[10]

Since the permanent income elasticity of consumption is unity, the consumption and permanent income elasticities of desired velocity are equal. To begin, the consumption regressions imply the following desired-velocity equations:

$(E/M)^*$ = constant + (coefficients of)	$t\times 10^{-2}$	(E/P)
E2′	1.566	−.806
E3′	1.135	−1.053
E4′	.585	−.834

We emphasize that (a) these equations are estimated only indirectly,[11] and (b) they are derived from data taken entirely from a period in which the long-established velocity-income relationship seems to have been reversed completely. In spite of this, both estimates are consistent with that long-run relationship and with the behavior of postwar velocity.[12]

If interest rate movements alone could account for the postwar rise in velocity, we should have found significantly positive interest rate coefficients and insignificant constant terms in the Table 1 regressions. In fact, we find just the opposite. This implies that those who have explained the postwar rise in velocity by interest rate movements have confounded interest rates with some other trend-related phenomena.

If interest rates cannot then account for any part of the postwar rise in velocity, some part might be attributable to a change in the relationship between velocity and income. In that event, we should find some combination of positive secular trend and an absolute decline in the income elasticity of velocity. However, no evidence that the long-run income-velocity relationship has been altered is found. Indeed, it is remarkable how close my results come to duplicating exactly that long-run relationship. Friedman [3, p. 338] has estimated directly a money-demand equation very close in form to mine using cycle-average data from 1869–1957. It contains a permanent income elasticity of velocity of −.810. Two of my three estimates of this parameter are essentially identical to this.[13] Laidler's estimate of

[10] The broader money definitions seem to work slightly better (in particular E3 is the best consumption regression and E7 the best net national product regression), but this may be an aberration of the period covered by our data. There were several changes in the maximum interest rate payable on time deposits of commercial banks, and this led to some once-and-for-all shifting from demand to time deposits. This source of error in the conventional money series is reduced somewhat by the use of first differences in estimating the regressions.

[11] They are derived by dividing the constant term and the coefficient of the real expenditures variable by the absolute value of the coefficient of the velocity variable (see equations (1) through (4″)), and then expressing the result in levels rather than first differences.

[12] When we plug the 1953–1966 changes of the independent variables into the desired velocity equations we obtain the following comparison of actual and predicted (i.e., desired) changes:

Parameters Derived from and Velocity as Defined in	Percentage Change, 1953–1966: Desired Velocity	Actual Velocity
E2	49.56	53.15
E3	5.37	7.28
E4	−12.96	−12.08
E5	46.22	51.22
E6	3.60	5.93
E7	−14.58	−13.52

This close agreement is largely a by-product of the fitting procedure. The mean change in actual velocity will be given by the desired velocity equation evaluated at the means of changes in its independent variables, if the dependent variable in Table 1 has a zero mean. The mean of that variable is, in fact, close to zero.

[13] Friedman's money definition is that in E3′ where my estimate actually exceeds his in absolute value. Another aspect of Friedman's study that is interesting in light of my results is his discussion of the cyclical behavior of velocity. Friedman estimated his desired

−.618 for this elasticity [7, p. 546], derived from cycle-average data, 1891–1957, is in less spectacular, though close, agreement with my estimates.[14] My estimates, moreover, did not require cycle-averaging of the data.

The implicit income elasticity of velocity in the net national product regressions is never significantly negative and always close to zero in magnitude. This result seems, at first glance, to conflict with the conclusion that the long-run velocity-income relationship has held in the postwar period. However, essentially zero measured income elasticities of velocity have been reported for other data which extend over longer periods.[15] Measured income is simply a poor proxy for permanent income in any time period. In any event, the net national product regressions are even less congenial to an interest rate explanation of postwar velocity than the consumption regressions. All the interest rate coefficients here are negative, though insignificant.

In sum, these results mean that we can rely neither on interest rates nor on some change in long-established velocity-income relationships to explain the postwar rise in velocity. Some other forces, for which secular trend is here a proxy, are entirely responsible for this rise. This only brings the general problem into sharper focus. The important task for future research is to identify and measure the particular forces here subsumed in "trend." When this is done, and the trend term in desired velocity is thus rendered insignificant, our ability to predict the demand for money should be enhanced considerably.[16]

velocity equation by regressing permanent income on *actual* velocity. He found that the residuals from this relationship conformed negatively with and led the business cycle. My model explains this cyclical pattern: Friedman's "residual element" is the difference between actual and desired velocity, and this difference generates opposite-signed expenditure changes with a one quarter lag.

[14] Laidler estimates the permanent income-elasticity of money; this may be converted to the income-elasticity of velocity by subtracting it from one.

[15] Chow [1, p. 124] finds a measured income-elasticity of money of +.9298 for 1897–1958 data, which implies an income-elasticity of velocity of +.0702.

[16] Note that the constant terms in Table 1, from which the trend in desired velocity is deduced, get smaller as the definition of the money total is broadened. At the same time, the income coefficient is insensitive to the definition of the money total. This means that the income-elasticities of conventional money and its closest substitutes are about the same, but that some of the secular decline in the demand for conventional money shows up as a secular rise in the demand for bank and other financial institution time deposits. This result serves to emphasize the general problem raised in this paper, namely that the forces causing the secular shift from money to other goods and assets remain to be identified in future research.

REFERENCES

1. G. CHOW, "On the Long-Run and Short-Run Demand for Money," *Jour. Pol. Econ.*, April 1966, *74*, 111–31.
2. M. FRIEDMAN, "Interest Rates and the Demand for Money," *Jour. Law Econ.*, Oct. 1966, *9*, 71–85.
3. ———, "The Demand for Money: Some Theoretical and Empirical Results," *Jour. Pol. Econ.*, June 1959, *67*, 327–51.
4. ——— AND D. MEISELMAN, "The Relative Stability of Monetary Velocity and the Investment Multiplier in the United States," in Commission on Money and Credit, *Stabilization Policies*, Englewood Cliffs 1963.
5. ——— AND A. J. SCHWARTZ, *A Monetary History of the United States 1867–1960.* Princeton 1963.
6. M. HAMBURGER, "The Demand for Money by Households, Money Substitutes and Monetary Policy," *Jour. Pol. Econ.*, Dec 1966, *74*, 600–623.
7. D. LAIDLER, "The Rate of Interest and the Demand for Money," *Jour. Pol. Econ.*, Dec. 1966, *74*, 543–55.
8. ———, "Some Evidence on the Demand for Money," *Jour. Pol. Econ.*, Feb. 1966, *74*, 55–68.

A Model of Labor Migration and Urban Unemployment in Less Developed Countries

By MICHAEL P. TODARO*

The chronic problem of urban unemployment and underemployment in almost every contemporary developing country has received a relatively minimal degree of theoretical attention in the literature on economic development. However, even the most casual observer of these countries cannot help but be overwhelmed by the proportion of the urban labor force which is apparently untouched by the so-called "modern" economy. From Dar es Salaam to Karachi to Caracas, from land surplus to labor surplus to capital surplus countries, one hears of the ever-increasing flow of rural migrants into urban areas and of the inability of the urban economy to provide permanent jobs for even a majority of these workers.[1] And yet, in striking contrast to the sophisticated theories of unemployment in developed nations, there have been few attempts to formulate a realistic positive theory of urban unemployment for less developed countries.[2] In fact, one of the best known models of labor transfer and economic development does not even consider the causes or, for that matter, the implications of a large and rapidly growing pool of urban unemployed [23].

The objective of this paper is twofold. First, we shall formulate an economic behavioral model of rural-urban migration which, in our opinion, represents a realistic modification and extension of the simple-wage differential approach commonly found in the literature [16] [19] [23]. It does so by recognizing the fact that the existence of a large pool of unemployed and underemployed urban workers must certainly affect a prospective migrant's "probability" of finding a job in the modern sector. As a result, when analyzing the determinants of urban labor supplies, one must look not at prevailing real income differentials as such but rather at the rural-urban "expected" income differential, i.e., the income differential adjusted for the probability of finding an urban job. It will be argued that this probability variable acts as an equilibrating force on urban unemployment rates. Secondly, we shall incorporate this probabilistic approach into a rigorous model of the determinants of urban labor demand and supply which,

* The author is research fellow at the Institute for Development Studies in Nairobi, Kenya. He wishes to thank Emile Despres, Peter Diamond, John Harris, Lloyd Reynolds, Joseph Stiglitz, and the referee of this journal for their very helpful comments and suggestions on earlier drafts of the paper. Naturally, he alone bears responsibility for any remaining defects.

[1] W. A. Lewis stands out among development economists as one who has repeatedly called our attention to the seriousness of the urban unemployment problem [21] [22]. However, Lewis' discussions have been largely qualitative and have not provided any rigorous framework with which to analyze the mechanism of labor migration and urban unemployment.

[2] Eckaus' famous factor proportions model [6] represents the most notable attempt to come to grips in a rigorous fashion with the problem of labor absorption in the modern sector. However, his model is concerned primarily with the demand side of the employment problem, and as such does not consider in an equally rigorous fashion the determinants of rural-urban labor supply. As a result, the model cannot be used to estimate the magnitude of urban unemployment nor can it be used to evaluate the unemployment implications of alternative policies.

when given values for the crucial parameters, can be used among other things to estimate the equilibrium proportion of the urban labor force that is not absorbed by the modern industrial economy. Additionally, the model will provide a convenient framework for analyzing the implications of alternative policies designed to alleviate unemployment by varying one or more of the principal parameters.

I. *The Process of Labor Migration*

It is a well-known fact of economic history that material progress usually has been associated with the gradual but continuous transfer of economic agents from rural based traditional agriculture to urban oriented modern industry [5] [27]. It is not surprising, therefore, to find the literature on economic development stressing the importance of similar structural changes in contemporary less developed nations [4] [20]. In particular, with respect to the occupational distribution of the indigenous labor force, economic development is often defined in terms of the transfer of a large proportion of workers from agricultural to industrial activities [24]. However, this process of labor transfer is typically viewed analytically as a one-stage phenomenon, that is, a worker migrates from a low productivity rural job directly to a higher productivity urban industrial job. The question is rarely asked as to whether or not the typical unskilled rural migrant can indeed find higher-paying regular urban employment. The empirical fact of widespread and chronic urban unemployment and underemployment attests to the implausibility of such a simple view of the migration process.

It is our opinion that a more realistic picture of labor migration in less developed countries would be one that views migration as a two-stage phenomenon. The first stage finds the unskilled rural worker migrating to an urban area and initially spending a certain period of time in the so-called "urban traditional" sector.[3] The second stage is reached with the eventual attainment of a more permanent modern sector job. This two-stage process permits us to ask some fundamentally important questions regarding the decision to migrate, the proportionate size of the urban traditional sector, and the implications of accelerated industrial growth and/or alternative rural-urban real income differentials on labor participation in the modern economy.

II. *Employment Probability and the Decision to Migrate*

In our model, the decision to migrate from rural to urban areas will be functionally related to two principal variables: (1) the urban-rural real income differential and (2) the probability of obtaining an urban job. Since it is this latter variable which will play a pivotal role in the analysis, it might be instructive at this point to explain briefly our reasons for incorporating this probability notion into the overall framework.

As pointed out above, an implicit assumption of typical labor transfer models is that any migrant who enters the modern sector is "absorbed" into the gainfully employed at the prevailing urban real wage. However, the important question to ask in this context is "how long" does the

[3] For the purposes of this paper, the urban traditional sector will encompass all those workers not regularly employed in the urban modern sector, i.e., the overtly unemployed, the underemployed or sporadically employed, and those who grind out a meagre existence in petty retail trades and services. J. P. Lewis provides an excellent description of this traditional sector which consists largely of "the urban in-migrant who, instead of doing absolutely nothing, joins Bombay's army of underemployed bootblacks or Delhi's throngs of self-appointed (and tippable) parking directors, or who becomes an extra, redundant salesman in the yard goods stall of the cousin, who according to custom, is going to have to provide him with bed and board anyway" [18, p. 53]. This description aptly fits a typical city in Africa and Latin America as well.

average migrant have to wait before actually obtaining a job. Even if the prevailing real wage is significantly higher than expected rural income, the fact that the "probability" of obtaining a modern sector job, say within the next year or two, is very low must certainly influence the prospective migrant's choice as to whether or not he should leave the farm. In effect, he must balance the probabilities and risks of being unemployed or sporadically employed in the city for a certain period of time against the favorable urban wage differential. A 70 per cent urban real wage premium, for example, might be of little consequence to the prospective migrant if his chances of actually securing a job are, say, one in fifty. Nevertheless, even if expected urban real income is less than rural real income for a certain period following migration, it may still be economically rational from a longer-run point of view (e.g., from a discounted present value approach to the rural-urban work choice) for the individual to migrate and swell the ranks of the urban traditional sector. Our underlying behavioral model, therefore, will be formulated more in the spirit of permanent income theories than present wage differential theories.

To underline the fundamental role played by job opportunities and probabilities of employment in the actual migration decision-making process, we might cite two outstanding illustrations, one historical and one contemporary, which demonstrate the relative, and often overriding, importance of this variable. The first case concerns the movements of American unskilled laborers back and forth between agriculture and industry during the 1930 depression decade. In an extremely informative and well-documented study of American agriculture, Theodore Schultz [26] argues that in 1932 when urban wages were still considerably higher and falling less rapidly than rural wages, there was a definite reversal of the historical flow of workers from the farm to the city. In fact, 1932 witnessed a net urban to rural labor migration [26, p. 90]. Schultz attributes this seemingly paradoxical phenomenon to the severe lack of job opportunities in depressed urban factories and the more likely prospects of finding agricultural employment in rural areas even though there still existed a significant positive urban wage premium [26, p. 99].

The second, more contemporary case concerns an interesting experiment carried out in Kenya in 1964. In a modified version of a tactic suggested by the International Labor Office [29] which advocated that governments of less developed countries employ and, through taxes and subsidies, induce private enterprise to employ more labor than would be worthwhile on the basis of a comparison between productivity and wages, the government of Kenya instituted a "tripartite agreement" among itself, private employers, and trade unions. The avowed intention was to wipe out the considerable unemployment existing in the greater Nairobi area by having the two hiring participants agree to increase their employment immediately by 15 per cent. For their part the unions had to agree to forego any demands for general wage increases. In his analysis of this "agreement" Professor Harbison has observed that:

> The effort was a colossal failure. The private employers did take on additional workers and *this acted like a magnet attracting new workers into the urban labor markets;* in a few months the working forces in most of the private establishments had dropped to their former levels through attrition not offset by new hires. In the end, the volume of unemployment, as a consequence of the expansion of the modern labor force *in response to the prospect of more jobs* was probably increased rather than de-

creased [10, p. 183, fn**]. (Italics not in original)

Here once again we can recognize the basic influence exerted by the probability of finding a job (whether real or anticipated) on the supply of rural workers into urban labor markets. Moreover, the significance of anticipated job opportunities on urban labor supplies is underlined in the above case by the enforced stability of urban wages over this experimental period. This tripartite agreement, therefore, seems to have provided as excellent a *ceteris paribus* experiment as is normally possible in economic analysis. In terms of the model that we shall now present, the Kenyan experience would be interpreted as a rightward shift of the urban labor supply curve as a result of an anticipated increase in the probability of successfully locating a job in the modern sector.

III. *A Behavioral Model of Rural-Urban Labor Migration*

In order to understand better the nature of the supply function to be used later in the overall model of the determinants of urban unemployment, let us first set forth the underlying behavioral assumptions of our model of rural-urban migration.

1. We shall assume that the percentage change in the urban labor force as a result of migration during any period is governed by the differential between the discounted streams of expected urban and rural real income (defined below) expressed as a percentage of the discounted stream of expected rural real income[4]—i.e.,

$$(1)\quad \frac{\dot{S}}{S}(t) = F\left[\frac{V_u(t) - V_R(t)}{V_R(t)}\right], \quad F' > 0.$$

where,

$\dot{S}$ represents net rural urban migration;

S is the existing size of the urban labor force;

$V_u(t)$ is the discounted present value of the expected urban real income stream over an unskilled worker's planning horizon; and,

$V_R(t)$ is the discounted present value of the expected rural real income stream over the same planning horizon.

2. The planning horizon for each worker is identical.

3. The fixed costs of migration are identical for all workers.

4. The discount factor is constant over the planning horizon and identical for all potential migrants.

Given these initial assumptions, our behavioral urban labor supply model can be formulated in the following manner. First, for $V_R(0)$, we have:

$$(2)\qquad V_R(0) = \int_{t=0}^{n} Y_R(t)e^{-rt}dt$$

where,

$Y_R(t)$ represents net expected rural real income in period t based, say, on the average real income of x previous periods,[5] and

r is the discount factor reflecting

[4] As the reader will discover below when the overall model is presented, the reason for expressing migration as a percentage of the existing urban labor force rather than as an absolute number is that it greatly simplifies the mathematics without in any way altering the qualitative nature of our conclusions. Since we are interested in changes in unemployment rates over time, the crucial supply variable is the rate of increase in the urban labor force as a result of migration. The fact that equation (1) indicates that the same capitalized percentage earnings differential might lead to different absolute numbers of migrants at different times is reasonable as long as (a) the geographic distribution of the total population is heavily rural based and (b) the natural rate of rural population growth exceeds that of urban population growth. Both of these assumptions are generally valid for less developed nations.

[5] Thus, for example, one might anticipate expected rural incomes to decrease following periods of crop failure due to weather and pest variations. In our model, this would show up as an initial increase in migration equilibrated by a lower probability of finding an urban job which in turn lowers expected urban incomes as well.

the degree of consumption time preference of the typical rural unskilled worker.

Next, for $V_u(0)$, we have:

$$(3)\quad V_u(0) = \int_{t=0}^{n} p(t)\,Y_u(t)e^{-rt}dt - C(0)$$

where,

$Y_u(t)$ represents net urban real income in period t,[6]

$C(0)$ is the initial fixed cost of migration and relocation in the urban area,

and,

$p(t)$ is the probability of having a modern sector job in period t [defined below].

The distinguishing characteristic of equation (3) is that "expected" urban real income in any period t varies directly with $p(t)$, the probability of having a job in that period.[7] Thus, one could easily conceive of a situation in which the urban-rural real income differential, $Y_u(t)-Y_R(t)$, was positive while the "expected" differential, $p(t)Y_u(t)-Y_R(t)$, was negative.

Let us now consider the nature of $p(t)$. However, in order to give $p(t)$ a precise and intuitively plausible definition, it is necessary to look once again at the urban labor market and, in particular, the migration process. For analytical convenience, we shall picture the typical rural migrant, therefore, as arriving in the urban area and joining a large pool of unemployed and underemployed workers who arrived in town earlier and still are waiting for a modern sector job. The selection from this pool in each period is assumed to be random with the probability of selection being equal to the ratio of new job openings relative to the number of workers in this urban traditional sector. Since the probability of *having* a job in any period, $p(t)$, is directly related to the probability of having been *selected* from the pool of urban traditional workers in that or any previous period, we can formulate the relationship between these two variables in the following way. Let, $\pi(t)$ be the probability of being selected from the pool of urban traditional workers during period t if the worker is a member of that pool in period t; and let $p(t)$ be, as before, the probability of *having* a job in the urban modern sector in period t. It follows that,

$$p(0) = \pi(0)$$

and that,

$$p(1) = \pi(0) + (1 - \pi(0))\pi(1)$$

that is, the probability of having a job in period zero (the time of migration) is equal to the probability of immediate selection from this pool, while the probability of having a job in period 1 is equal to the probability of being selected in period zero *plus* the probability of being selected in period 1. Generalizing, we see that for any period, t,

$$p(t) = p(t-1) + [1 - p(t-1)]\pi(t)$$

or,

$$(4)\quad p(t) = \pi(0) + \sum_{i=1}^{t} \pi(i) \prod_{j=0}^{i-1} (1 - \pi(j))$$

where

$$\prod_{i=1}^{n} a_i = a_1 \cdot a_2 \cdot a_3 \cdot a_4 \cdots a_{n-1} \cdot a_n.^{8}$$

[6] $Y_u(t)$ is intended as a proxy variable for all elements that constitute urban real income, i.e., wages, cost of living, urban amenities etc. For example, Y_u might equal aw/p where w is the urban wage, p is the urban price deflator, and $a>1$ is scalar proxy for "city lights" and other amenities.

[7] Our reason for not including a similar probability variable in (2) is that the existence of traditional crop-sharing activities and the so-called "extended family" system largely negates the potential impact of such a variable in the rural economy whereas these ties are much more difficult to maintain in a wage-oriented urban economy. However, it would not be difficult to incorporate, say $P_R(t)$ into equation (2).

[8] One of the benefits of formulating the job selection process in this manner is that it captures an essential feature of the earnings history of a typical migrant,

In order to complete our behavioral model, we must now define $\pi(t)$ in some meaningful economic sense. We shall define this probability of being selected for a job during period t as being equal to the ratio of new modern sector employment openings in period t relative to the number of accumulated job seekers in the urban traditional sector at time t.[9] But this procedure necessitates the introduction of a demand expression to reflect job creation in the modern sector. For the purposes of this paper, therefore, we shall assume initially that the number of new jobs created increases at a constant exponential rate over time. Specifically

$$N(t) = N_0 e^{(\lambda-\rho)t} \tag{5}$$

where

- $N(t)$ is total modern sector employment in period t,
- λ is the rate of industrial output growth, and
- ρ is the rate of labor productivity growth in the modern sector.

Thus, if we let the rate of job creation $\gamma=\lambda-\rho$, we have

$$\pi(t) = \frac{\gamma N(t)}{S(t) - N(t)} \tag{6}$$

IV. *An Analytical Model of the Structure and Mechanism of Urban Labor Markets*

We can now bring some of the above concepts together and formulate our overall model of disturbances and adjustments in the urban labor markets of less developed countries. The model can be specified in the following manner. Once again, let

- $N(t)$ total employment in the urban modern sector in period t,
- $S(t)$ total urban labor force in period t

and, therefore,

$S(t)-N(t)$ measures the size of the urban traditional sector. We first have our exponential demand equation,

$$N(t) = N_0 e^{\gamma t}, \quad \text{or} \tag{7}$$

$$\frac{\dot{N}}{N}(t) = \gamma \tag{7a}$$

where,

$$\gamma = \lambda - \rho.$$

Next, we specify an aggregate labor supply equation which is a simplified version of equation (1) in the sense that only a one-period time horizon is assumed.[10]

$$\frac{\dot{S}}{S}(t) = \beta + \pi(t)F\left[\frac{Y_u(t) - Y_R(t)}{Y_R(t)}\right] \tag{8}$$

or, letting

$$\alpha(t) = \frac{Y_u(t) - Y_R(t)}{Y_R(t)}.$$

$$\frac{\dot{S}}{S}(t) = \beta + \pi(t)F(\alpha(t)) \tag{8a}$$

where,

- β is the natural rate of increase of the urban labor force,
- $\alpha(t)$ is the percentage urban-rural real income differential, and, is
- $F(\alpha(t))$ a function that such $dF/d\alpha>0$.

Thus, $\pi(t)F(\alpha(t))$ is the rate of urban labor force increase as a result of migration, i.e.,

namely, that the path of expected urban earnings is positively related to the length of time that a migrant has been in the urban area *ceteris paribus*. The longer a migrant remains in the urban area the more contacts he can establish and the more likely he is to be holding a job after a certain period of time. In terms of equation (4) $p(t)\rightarrow 1.0$ as $t\rightarrow\infty$.

[9] Since there will be other new migrants entering the labor pool during period t, the actual *realized* probability will be somewhat less than the expected probability at the time of choice (the latter being the more relevant criterion for migration). However, this slight differential will not affect our results or conclusions.

[10] This assumption is made necessary by mathematical convenience but is in fact probably a more realistic formulation in terms of actual decision making in less developed nations. In any case, the general conclusions are not sensitive to the assumption.

we are assuming that migration varies directly with the probability of finding a job. Furthermore, from equation (6), we know that

$$\pi(t) = \frac{\gamma N(t)}{S(t) - N(t)} \cdot$$

Substituting for $\pi(t)$ we have, therefore,

$$\text{(9)} \qquad \frac{\dot{S}}{S}(t) = \beta + \frac{\gamma N(t)}{S(t) - N(t)} (F\alpha(t)) \cdot$$

We shall assume, initially, that this income differential $\alpha(t)$ remains constant over time, i.e. $\alpha(t) = \alpha$.

Finally, we denote the proportion of the urban labor force employed in the modern sector at time t as $E(t)$, where

$$\text{(10)} \qquad E(t) = \frac{N(t)}{S(t)} \cdot$$

Before solving for equilibrium conditions, let us first give a brief verbal explanation of the mechanics of the model represented by equations (7a), (9), and (10).

Suppose we consider a developing economy in the very early stages of industrialization such that almost the entire population resides in rural areas. The urbanization process is just beginning to accelerate but as yet the pool of urban unemployed is relatively small so that the probability of obtaining a job is high. Therefore, given a significantly positive urban real wage premium $(\alpha > 0)$[11] and a positive rate of urban job creation exceeding the natural rate of urban population growth $(\gamma > \beta)$, the resulting urban expected real income differential induces rural-urban migration such that the urban labor force grows at a faster rate than that of job creation—i.e., $\beta + \pi(t)F(\alpha) > \gamma$. This more rapid growth of labor supply results in an increase in the relative size of the urban traditional sector with the result that *ceteris paribus* the probability of a rural migrant finding a job in the next period is somewhat lower $(\pi(t+1) < \pi(t))$. Assuming α and γ remain constant, this lower probability should result in a slowing down of the rate of urban labor force growth although $\dot{S}/S$ may continue to exceed $\dot{N}/N$. Eventually, however, the equilibrating function of π stabilizes the urban unemployment rate at some level $1-E^*$ depending upon the values of α, β, and γ. If the unemployment rate falls below $1-E^*$ equilibrating forces in the form of rising π's will be set in motion to restore the equilibrium. Thus, for any given values for our principal parameters, the equilibrium will be stable. Moreover, as we shall discover below, policies designed to eliminate unemployment by raising γ (e.g., by increasing the rate of industrial expansion and/or subsidizing labour in accordance with shadow price criteria) without a concentrated simultaneous effort at lowering the real earnings differential α will meet with increasing frustration. But now let us turn to a more rigorous demonstration of these and other conclusions.

The equilibrium condition for our model is defined simply as that employment rate

[11] Considerable debate has been generated in the literature as to whether this wage differential is in fact a real income differential. Lewis has argued that a positive money wage differential of the order of 50 per cent is necessary to induce workers to migrate from rural to urban areas [19] [22]. But this does not necessarily imply a corresponding real wage differential. Hagen [8] has argued and provided empirical evidence that this differential is in fact a distorted real-income differential resulting from disproportionate growth rates in manufacturing versus agricultural activities. See also comment by Koo [17], Hagen's reply [9], and remarks by Bhagwati and Ramaswami [2]. Recent empirical evidence has tended to confirm the real-income differential hypothesis [25] [13] [1] [12] [28] and Lewis himself has recognized the fact that in numerous cases unskilled workers in the modern sector are earning three and four times as much as the average small farmer [21, p. 12]. He attributes this disproportionate disequilibrium to the combined effects of trade union pressure, nationalistic government sympathy for the trade union cause, and a new social conscience on the part of big entrepreneurs. Whatever the reasons, these real-earnings distortions are a major cause of the urban unemployment phenomenon.

E^* such that $\dot{E}/E(t)$ equals zero, that is, where

$$\frac{\dot{E}}{E}(t) = \frac{\dot{N}}{N}(t) - \frac{\dot{S}}{S}(t) = 0. \tag{11}$$

Now, from equations (7a) and (9) we know that

$$\frac{\dot{E}}{E}(t) = \gamma - \beta - \frac{\gamma F(\alpha)N(t)}{S(t) - N(t)} = 0$$

where,

$$\gamma - \beta = \frac{\gamma F(\alpha)N(t)}{S(t) - N(t)}. \tag{12}$$

Dividing both numerator and denominator of the right-side term of (12) by $S(t)$ and substituting from (10) we obtain:

$$\gamma - \beta = \frac{\gamma F(\alpha)E^*}{1 - E^*}. \tag{13}$$

Rearranging,

$$\gamma - \beta - (\gamma - \beta)E^* = \gamma F(\alpha)E^*$$

or, finally:

$$E^* = \frac{\gamma - \beta}{\gamma F(\alpha) + \gamma - \beta}. \tag{14}$$

Alternatively, the equilibrium proportionate size of the urban traditional sector, $T^* = 1 - E^*$, is simply:

$$T^* = 1 - \frac{\gamma - \beta}{\gamma F(\alpha) + \gamma - \beta}. \tag{15}$$

Furthermore, this is a stable equilibrium since from

$$\frac{\dot{E}}{E} = -\frac{\gamma F(\alpha)E}{1 - E} + \gamma - \beta = 0$$

we may show the derivative in (16) to be negative.

$$\frac{d\left(\frac{\dot{E}}{E}\right)}{dE} = -\frac{(1 - E)\gamma F(\alpha) + \gamma F(\alpha)E}{(1 - E)^2} \tag{16}$$

where

$$0 \leq E \leq 1.$$

It is evident from equation (15) that the proportionate equilibrium size of the urban traditional sector (T^*) will vary directly with the urban rural percentage real income differential, $(\partial T^*/\partial\alpha > 0)$, and inversely with the rate of job creation $(\partial T^*/\partial\gamma < 0)$. Moreover, it is interesting to note that an increase in the rate of industrial output growth (λ) which in turn increases the growth rate of modern sector employment opportunities (γ) might have no impact on cutting into the proportionate size of the urban traditional sector if the urban real wage differential (α) also increases by a certain amount. Specifically, $dE^* = 0$, if, equation (17) holds,

$$\frac{\gamma + d\gamma - \beta}{F(\alpha)\gamma + \gamma dF(\alpha) + F(\alpha)d\gamma + dF(\alpha)d\gamma + \gamma + d\gamma - \beta} = \frac{\gamma - \beta}{F(\alpha)\gamma + \gamma - \beta} \tag{17}$$

or, solving for $d\gamma$, if

$$d\gamma = \frac{-\gamma^2 dF(\alpha)}{\gamma dF(\alpha) - \gamma\beta - F(\alpha)\beta - \beta dF(\alpha)}. \tag{18}$$

For example, suppose the growth rate of modern sector employment is 4 per cent ($\gamma = .04$), the natural rate of urban labor force growth is 2 per cent ($\beta = .02$), the urban-rural real earnings differential is 100 per cent ($\alpha = 1.0$), and, for simplicity $F(\alpha) = \alpha$: Given these parameters equation (14) says that, in equilibrium, modern sector employment would absorb only one-third of the urban labor force. Now suppose that the earnings differential between modern urban jobs and traditional agricultural work increases by an additional

20 per cent, i.e., $d\alpha = dF(\alpha) = 0.20$. Equation (18) of our model says that the rate of modern sector job creation must grow by an additional 1.9 per cent (i.e., $d\gamma = .019$) just to prevent the equilibrium employment rate from falling below its original level. Moreover, when it is recalled that $\gamma = \lambda - \rho$ and we recognize the fact that in order to increase employment growth (γ) by 2 per cent, the growth rate of modern sector output (λ) will probably have to increase by at least an additional 6 per cent due to the positive corollation between output expansion and productivity growth,[12] we begin to appreciate the great difficulty of absorbing larger proportions of the urban labor force without a concentrated effort designed to prevent the further widening of urban-rural real earnings differentials.

Perhaps a more interesting and relevant application of equation (18) is to consider the potentially conflicting objectives of a successful program of import substitution and a concomitant reduction of modern sector unemployment rates. Bruton has recently underlined the necessity for productivity to grow if import substituting industries are to pay for themselves in real terms [3]. Consequently, it is extremely important that labor productivity should increase substantially in the modern sector. But, as the above example demonstrates, if the gap between urban and rural real earnings capacity is permitted to widen further, the likelihood of simultaneously raising labor productivity and lowering urban unemployment rates appears negligible indeed. However, if in the above example the urban-rural income differential were to contract by 20 per cent, equation (18) tells us that labor productivity could expand by an additional 1.3 per cent annually without increasing the urban unemployment rate. Alternatively, labor productivity could expand by, say, an additional 1 per cent per annum with a simultaneous decline in the urban unemployment rate.

Finally, consider the question of agricultural development strategy. Johnston has strongly emphasized the point that if the agricultural sector is to make its most meaningful contribution to economic development, it must not only improve labor productivity but also expand employment opportunities [14] [15]. The main point is that premature mechanization of agriculture through the adoption of the most modern techniques of large-scale farming poses serious problems for rural labor absorption. In terms of our model, Johnston's argument would indicate a lowering of the expected rural wage through the introduction of a probability variable similar to that in the urban sector and *ceteris paribus* a consequent rise in rural-urban labor migration. The implication here is that if employment creation is high on the priority list of developing countries, not only should the real wage differential be prohibited from increasing through some appropriate incomes policy but also output and productivity growth in agriculture wherever feasible must be achieved through more efficient use of existing capital resources and not through capital-labor substitution.[13]

[12] For some cross-sectional statistical evidence of the positive correlation between rates of output growth and rates of labor productivity growth in the manufacturing sectors of less developed nations (and hence the employment lag), see United Nations [30, p. 96–98]. The principal explanations offered for the observed rapid increases in labor productivity include the following: a greater substitution of capital for labor than is warranted by shadow prices of labor and capital [22, pp. 55–68], improved managerial and organizational capabilities [25, p. 33], and an upgrading of labor efficiency through on the job training programs and the emergence of a more stable, proletarian urban labor force [7].

[13] Numerous other important issues of development policy as they relate to the urban unemployment problem take on a new and often surprising aspect when

V. *Some Concluding Comments and Implications of the Analysis*

We have attempted in this paper to formulate a model that is both descriptive and analytical with respect to the mechanism through which economic variables influence urban labor markets in less developed countries. The analysis has tried to come to grips in a simple but rigorous fashion with the fundamental factors affecting urban unemployment and underemployment. Although the model has made some simplifying assumptions especially with respect to the constancy of γ and α, the overall net impact of allowing these parameters to vary over time and/or choosing alternative values is demonstrated, in our opinion, quite conveniently in the model. Moreover, it underlines in a simple and plausible way the interdependent effects of industrial expansion, productivity growth, and the differential expected real earnings capacity of urban versus rural activities on the size and rate of increase in labor migration and, therefore, ultimately on the occupational distribution of the urban labor force.

Perhaps the most significant policy implication emerging from the model is the great difficulty of substantially reducing the size of the urban traditional sector without a concentrated effort at making rural life more attractive. For example, instead of allocating scarce capital funds to urban low cost housing projects which would effectively raise urban real incomes and might therefore lead to a worsening of the housing problem, governments in less developed countries might do better if they devoted these funds to the improvement of rural amenities. In effect, the net benefit of bringing "city lights" to the countryside might greatly exceed whatever net benefit might be derived from luring more peasants to the city by increasing the attractiveness of urban living conditions. Like Marshall's famous scissors analogy, the equilibrium level of nonparticipation in the urban economy is as much a function of rural "supply push" as it is one of urban "demand pull." Thus, as long as the urban-rural real income differential continues to rise sufficiently fast to offset any sustained increase in the rate of job creation, then even in spite of the long-run stabilizing effect of a lower probability of successfully finding modern sector employment, the lure of relatively higher permanent incomes will continue to attract a steady stream of rural migrants into the ever more congested urban slums. The potential social, political, and economic ramifications of this growing mass of urban unemployed should not be taken lightly.

analyzed within the framework put forth in this paper. For example, in a forthcoming paper which incorporates our basic migration model into a more general two-sector model of economic development, it is demonstrated that under certain (entirely plausible) conditions the economist's standard theoretical policy prescription of generating urban employment opportunities through the use of "shadow prices" implemented by means of wage subsidies or direct government hiring might in fact exacerbate the problem of urban unemployment [11]. Also, the pricing policies of agricultural marketing boards and the burden of rural taxation are revealed as important factors influencing the magnitude of urban unemployment while standard formulas for investment criteria are shown to be biased against agricultural investment projects.

REFERENCES

1. E. J. BERG, "Wage Structure in Less Developed Countries," forthcoming in a symposium volume on *Wage Policy Issues in Economic Development* to be published for the International Institute for Labour Studies.
2. J. N. BHAGWATI AND V. K. RAMASWAMI, "Domestic Distortions, Tariffs, and the Theory of Optimum Subsidy," *Jour. Pol. Econ.*, Feb. 1963, *71*, 44–50.
3. H. J. BRUTON, "Import Substitution and Productivity Growth," *Jour. Develop. Stud.*, Apr. 1968, *4*, 306–26.
4. H. B. CHENERY, "Patterns of Industrial

Growth," *Am. Econ. Rev.*, Sept. 1960, *50*, 624–54.
5. C. Clark, *Conditions of Economic Progress*. London 1957.
6. R. S. Eckaus, "The Factor Proportions Problem in Underdeveloped Areas," *Am. Econ. Rev.*, Sept. 1955, *45*, 539–65.
7. W. Elkan, *Migrants and Proletarians*. London 1960.
8. E. E. Hagen, "An Economic Justification of Protection," *Quart. Jour. Econ.*, Nov. 1958, *72*, 496–514.
9. ———, "An Economic Justification of Protection: Reply," *Quart. Jour. Econ.*, Feb. 1961, *75*, 145–51.
10. F. Harbison, "The Generation of Employment in Newly Developing Countries," in J. Sheffield (ed.), *Education, Employment and Rural Development*, Nairobi 1967.
11. J. R. Harris and M. P. Todaro, "A Two Sector Model of Migration with Urban Unemployment in Developing Economies," (Working Paper, Department of Economics, Massachusetts Institute of Technology, No. 33, Dec. 1968, mimeo)
12. J. K. Helleiner, "Agricultural Export Pricing Strategy in Tanzania," *East African Jour. Rural Develop.*, Jan. 1968, *1*, 1–18.
13. J. P. Henderson, Wage Policy in Africa," (mimeo), paper prepared for delivery at the African Conference on Economics, Temple University, April 1968, 1–20.
14. B. F. Johnston, "Agriculture and Economic Development: The Relevance of the Japanese Experience," *Food Research Inst. Stud.*, 1966, *6*, 251–312.
15. B. F. Johnston and S. T. Nielsen, "Agricultural and Structural Transformation in a Developing Economy," *Econ. Develop. and Cult. Change*, April 1966, *14*, 279–301.
16. D. W. Jorgenson, "The Development of the Dual Economy," *Econ. Jour.*, June 1961, *71*, 309–34.
17. A. Y. C. Koo, "An Economic Justification of Protection: Comment," *Quart. Jour. Econ.*, Feb. 1961, *75*, 133–44.
18. J. P. Lewis, *Quiet Crisis in India*. Washington 1962.
19. W. A. Lewis, "Economic Development with Unlimited Supplies of Labour," *Manchester School*, May 1954, *22*, 139–91.
20. ———, *The Theory of Economic Growth*. London 1955.
21. ———, "A Review of Economic Development," *Am. Econ. Rev.*, Proc., May 1965, *55*, 1–16.
22. ———, *Development Planning*. New York 1966.
23. G. Ranis and J. C. H. Fei, "A Theory of Economic Development," *Am. Econ. Rev.*, Sept. 1961, *51*, 533–56.
24. ———, *Development of the Labor-Surplus Economy: Theory and Policy*. Homewood 1964.
25. L. G. Reynolds, "Wages and Employment in a Labor-Surplus Economy," *Am. Econ. Rev.*, March 1965, *55*, 19–39.
26. T. W. Schultz, *Agriculture in an Unstable Economy*. New York 1945.
27. B. E. Supple, ed., *The Experience of Economic Growth*, New York 1963.
28. W. M. Warren, "Urban Real Wages and the Nigerian Trade Union Movement, 1939–1960," *Econ. Develop. and Cult. Change*, Oct. 1966, *15*, 21–36.
29. International Labor Office, *Employment and Economic Growth*. Geneva 1964.
30. United Nations, *The Growth of World Industry, 1938–61: International Analyses and Tables*. New York 1965.

Implications of Dynamic Monopoly Behavior

By Frank C. Jen and Lawrence Southwick, Jr.*

The behavior of a monopolist has traditionally been studied by models allowing only one seller to operate. With a few exceptions such as Stigler [12] and Brems [3] the models are also static in nature although further research in dynamic models is frequently suggested. The result is a lack of depth in the study of the effects of dynamic monopoly strategies such as limiting price and quantity. The question of how a seller initially protected by a patent (hereinafter a "protected monopoly") may behave over time cannot be answered either. However, a monopoly can be regarded as a member of a duopoly where the quantity produced by the other member is zero. Hence, a convenient way to study dynamic monopoly behavior is to construct a general dynamic duopoly model from which conditions for various types of monopoly can be derived. This paper constructs such a model and derives the "maximizing" equations in Section I. Section II derives the conditions under which a natural monopoly can exist. The possible production and price paths of a monopoly are also discussed. Section III studies the production and pricing strategies of a firm which has a monopoly protection initially and wants to continue the monopoly. Section IV compares the effects of collusion with those of monopoly and further studies the implications of this model with respect to oligopolistic behavior. Section V reviews the findings and discusses areas for futher research.

The traditional static monopoly model can be represented by the system of equations,[1] $P=f(X)$ and $C=g(X)$, where P, C, X are respectively unit price charged, total cost incurred, and quantity produced by the monopolist. The monopolist will then choose an X so that the necessary $MR=MC$ condition is satisfied:

$$(1)\qquad \frac{\partial Xf(X)}{\partial X}=\frac{\partial g(X)}{\partial X} \quad \text{or} \quad Xf_X+f=g_X$$

Implicit in the static approach is, however, an assumption that the monopolist has stable cost and demand curves over time. Moreover, it also implies that such existing barriers to entry as patent protection, cost advantages, etc., will last forever to protect the monopolist from competition. Both assumptions are questionable. Indeed, cost and demand curves may shift and change slope due to such factors as secular trend, technological advances, etc. Cost curves may also be significantly affected by costs incurred in changing the level of output over time as the pioneering study by Holt, *et al.* [9] has indicated. An adequate monopoly behavior model should therefore not only be dynamic in nature in order to account for possible future changes of cost and demand curves, but should also be inclusive enough so that the effect of possible future changes in

* Professor of finance and management science and assistant professor of management science respectively, State University of New York at Buffalo. The names of the authors are listed in alphabetical order. We benefited from discussions with our colleagues, particularly Ken Babka and Clark Strausser. Any remaining errors, of course, are ours.

[1] See, e.g. [5], [11], [12]. An assumption used in all models, e.g., [5, p. 188] is that the demand for the product must be reasonably independent of the price of other products.

monopoly position can be studied. The latter problem can be studied through a duopoly model because the existence of a monopoly with permanent protection from competition implies that the potential other producer in the model will always be facing barriers so effective that it will abstain from producing.[2] Hence, a dynamic duopoly model can be used to study the problem of dynamic monopoly behavior. Such a model is presented in the next section.

I. *A Dynamic Duopoly Model and its Solutions*

There is presumed to be one homogeneous good for which the demand curve is a function of time. At time t, firm 1 produces an amount X while firm 2 produces an amount Y. The price received by both firms is assumed to follow the simple deterministic but dynamic function:[3]

$$P = f(X, Y, t) = f \tag{2}$$

That is, for time t, after each firm determines its output, a market clearing price (P) is obtained. The function is symmetric in X and Y with negative first partial derivatives; $f_X = f_Y < 0$.

The total cost for each firm in the duopoly may, however, differ. For each firm, we assume that the total cost (g or g') is a function of : (1) its output level (X or Y), (2) the rate of change of its output level ($\dot{X}$ or $\dot{Y}$), and (3) time (t or $t-t_0$).

$$\begin{aligned} C(X) &= g(X, \dot{X}, t) = g \\ C(Y) &= g(Y, \dot{Y}, t - t_0) = g' \end{aligned} \tag{3}$$

The output level variable is introduced to account for the effects of scale on costs. The rate of change variable, which is not included in the static model, is introduced to account for the effects of changing output level on total costs. According to Holt, *et al.* [9], the profitability of a firm is also dependent on the time paths of the production level and, further, the effect of changing production level on cost is of U-shaped form centered around zero. Time is introduced in this model to account for possible increases or decreases in cost due to such factors as secular trend and technological change. By manipulating t, we can also allow one firm to have a lead. In this connection, because independent innovators will probably not start producing a new product at the same time, we assume that firm 1 has a time lead of t_0[4] during which firm 2 is not allowed to produce and further, *frm 2's cost function will always lag firm 1 by t_0 periods.*[5] (Firm 1 thus has a protected monopoly from period 0 to t_0.)

Note that we have chosen the same functional form for firms 1 and 2 for the cost equations in (3). The choice is made in order to make everything as much the same for both firms as possible so that the behavior of these two firms can be compared analytically. Note further that total cost can also be a function of interactions of second or higher orders between $X, \dot{X}, t$ and $Y, \dot{Y}, t$. Thus, while the model appears as very similar to the static model, the basic relationships can be far more complex.

Under the condition that each firm is to maximize the present value of its future

[2] If there are N possible entrants, the monopolist needs only to regard the most efficient one as the potential other producer in the duopoly model. Uncertainties however may increase drastically as N gets large so that the duopoly model will not apply.

[3] We use the simplest dynamic price function possible to make the mathematics manageable.

[4] For example, if a firm has patent protection, other firms are then not allowed to produce within the period of patent. In such cases, the period of patent protection will be t_0.

[5] This assumption will cause the late starter to suffer from a cost disadvantage, which may be important for maintaining the monopoly power of the first firm as demonstrated later in this paper. This assumption will not hold if the late starter also benefits from the first firm's effort to develop the product by creating consumer acceptance or improving technology.

profit stream, firm 1 chooses a *time path* of X to maximize its present value:

$$(4)\quad PV(1) = \int_0^{\infty} \exp\{-rt\}[Xf(X, Y, t) - g(X, \dot{X}, t)]dt$$

From t_0 on, firm 2 chooses a time path of Y to maximize:

$$(5)\quad PV(2) = \int_{t_0}^{\infty} \exp\{-rt\}[Yf(X, Y, t) - g(Y, \dot{Y}, t - t_0)]dt$$

The same cost of capital (r) has been used for both firms. This choice is made for convenience.[6]

Assuming continuity, a *necessary* condition for maximizing the present values of firms 1 and 2, *independently taken*, is the Euler condition.[7] For firms 1 and 2, the resulting equilibrium conditions are:

$$(6)\quad \frac{f + Xf_X + Xf_Y Y_X}{r} = \frac{g_X}{r} + g_X - \frac{g_{\dot{X}t}}{r} - \frac{g_{\dot{X}X}\dot{X}}{r} - \frac{g_{\dot{X}\dot{X}}\ddot{X}}{r}$$

$$(7)\quad \frac{f + Yf_Y + Yf_X X_Y}{r} = \frac{g'_Y}{r} + g'_{\dot{Y}} - \frac{g'_{\dot{Y}t}}{r} - \frac{g'_{\dot{Y}Y}\dot{Y}}{r} - \frac{g_{\dot{Y}\dot{Y}}\ddot{Y}}{r}$$

[6] r can be considered as different between firms so that the effect of capital constraint on monopoly power can be studied. Useful analytical results will be difficult to obtain because they depend on the exact forms of the cost and demand functions.

[7] The Euler condition states that a necessary condition for an extremum to exist to the functional $\int_0^{\infty} h(z,\dot{z}, t)dt$ is:

$$\frac{\partial h}{\partial Z} = \frac{d}{dt}\frac{\partial h}{\partial \dot{Z}}$$

Clearly, this only holds for a relative extremum and further some complicated second order conditions must be checked to decide whether the extremum found is a relative maximum or minimum. However, equations for the second order conditions generally cannot be solved analytically. Hence, we assume that at least one of the solutions of the necessary conditions is a relative maximum; that is, we assume the second order conditions for the existence of a maximum are met for at least one of the solutions. See [6, Chs. 2 & 3].

In these two maximizing equations, the subscripted variables are partial derivatives with respect to the subscripts and the dotted variables are partial derivatives with respect to t. Variables Y_X and X_Y are, of course, the conjectural variations.

The two equilibrium equations are simultaneous second-order partial differential equations with two variables. The time paths of both X and Y and hence the time path of price could in principle be determined if two boundary conditions for both X and Y and values of the conjectural variations are specified. As to their meaning, since equations (6) and (7) contain similar terms, we will interpret (6) only. The three terms on the left-hand side are components of present value of future marginal revenue (MR). The first two terms in the numerator are similar to those in the static model in (1), but, here, they represent permanent streams of change in MR. The third term represents the MR lost or gained permanently by firm 1 due to price change caused by firm 2's change in output in response to firm 1's change. The r in the denominator, of course, translates the permanent stream of MR into its corresponding present value. As to the five terms on the right-hand side, they are components of present value of future marginal cost (MC). The first term is similar to that in the static model except that, as in the case of MR, the term now represents permanent change in MC and thus should be discounted by r to get its corresponding present value. The second term, $g_{\dot{X}}$, is a one-shot effect on marginal cost caused by incurring the cost of changing levels at time t while the third term, $-g_{\dot{X}t}/r$, is the present value of the permanent change in MC caused by the expected shift in the cost of changing levels

over time. The fourth term, $-g_{\dot{X}X}\dot{X}/r$, is the product of $-g_{\dot{X}X}/r$ and $\dot{X}$. The former is the present value of the permanent benefit (cost) to the larger firm in terms of decreased (increased) marginal cost of changing size while the latter is $\partial X/\partial t$. The fifth term, the product of $-g_{\dot{X}\dot{X}}/r$ and $\ddot{X}$, represents the effect of changing the rate of change of production levels on permanent marginal cost. Note that the signs of the last four terms may vary from firm to firm because they depend on the actual form of the cost function the firm faces from the scale, change of scale, and time points of view.[8]

Reverting to the particular problem of solving the simultaneous differential equations, we note that the initial boundary conditions are known: $X(0)=0$, $Y(t_0)=0$. The conjectural variations X_Y and Y_X are indeterminate, hence, analytical solutions for this dynamic duopoly model do not exist.

For the dynamic monopoly problem, however, analytical solutions do exist because the values of the conjectural variations can be assumed to be zero. We will therefore use this model to derive the *necessary* though not sufficient conditions for different types of monopoly to exist. Two types of monopoly will be studied. They are: (1) Natural Monopoly and (2) Imposed Monopoly.

II. *Natural Monopoly*

We will introduce a precise definition of natural monopoly in the context of our model: A firm is defined to have a natural monopoly *for a period of time* when the firm is producing according to its profit maximization conditions while the output of a potential second firm is zero according to the profit maximization conditions of the second firm. In the model presented, firm 1 will thus be a natural monopolist for a period while firm 2 has an output of zero given both firms maximizing their present values. The conditions for this natural monopoly to exist are, of course, for both maximizing equations (6) and (7) to hold for the special case $Y=\dot{Y}=\ddot{Y}=0$ and $Y_X=X_Y=0$. More specifically, the conditions are:

$$(8)\quad \frac{f(X,0,t)+Xf_X}{r}=\frac{g_X}{r}+g_{\dot{X}}-\frac{g_{\dot{X}t}}{r}-\frac{g_{\dot{X}X}\dot{X}}{r}-\frac{g_{\dot{X}\dot{X}}\ddot{X}}{r}$$

$$(9)\quad \frac{f(X,0,t)}{r}\leq\frac{g_Y(0,0,t-t_0)}{r}+g_{\dot{Y}}(0,0,t-t_0)-\frac{g_{\dot{Y}t}(0,0,t-t_0)}{r}$$

Equation (8) is a one-variable differential equation the solution of which is the production and price paths of firm 1. That is, firm 1 sets the production path, and thus the price path, so that the MR on the left-hand side equals the MC on the right. Equation (9) indicates that the price so set must be less then or equal to the MC *of the potential competitor* which consists of three terms: the present value of permanent change in MC of production due to scale, the effect of incurring a one-shot cost of altering production levels, and the present value of the permanent change in MC caused by the expected shift in the cost of altering production levels.

Equations (8) and (9) can further be combined so that the relationship between the MC of the monopolist and its potential competitor can be seen more clearly:

$$(10)\quad [g_Y(0,0,t-t_0)-g_X(X,\dot{X},t)]+r[g_{\dot{Y}}(0,0,t-t_0)-g_{\dot{X}}(X,\dot{X},t)]$$

[8] For instance, if we assume that the cost of changing level is declining over time, the third term is positive; that is, the firm will incur an additional stream of MC because it decides to change level immediately rather than later.

$$- [g_{\dot{Y}t}(0, 0, t - t_0) - g_{\dot{X}t}(X, \dot{X}, t)]$$
$$- g_{\dot{X}X}\dot{X} - g_{\dot{X}\dot{X}}\ddot{X} + Xf_X \geq 0$$

The first term measures the differences in MC due to scale between the potential competitor and the monopolist. How positive the term is depends, of course, on the economy of scale the monopolist enjoys and how much effect on MC the difference in the rate of change of production will have. The term is more positive for an industry which has a greater increasing return to scale, because a faster declining MC implies a greater increasing return to scale.[9] The latter is traditionally considered as a very important barrier to entry.

The second term measures the difference in MC of changing production levels. Although firm 1 may start from a non-zero rate of change, the scale economy can be dominating again, that is, firm 1 can have a cost of changing levels lower than firm 2 because large firms can have a smaller cost of changing level on a per unit basis.

The signs of the rest of the terms except the last are indeterminate because they differ from firm to firm as explained in Section I. The last term Xf_X has, of course, a negative sign because $f_X<0$. One can however conclude that a very important factor for a natural monopoly to exist is, in the context of our model, the scale economy. However, existence of this factor is *neither necessary nor sufficient* to support a natural monopoly as terms other than the scale economy in equation (10) can sometimes dominate. An example of dominance by a firm having a cost advantage because it enters the market earlier is presented below.

Example A. Assume a linear demand function:

$$P = m - nX - nY \tag{A.1}$$

Further assume that the only production costs are those associated with changing production levels. These costs declining over time at a constant rate k, approach zero asymptotically:

$$g = \frac{m}{r+k}\dot{X}\exp\{-kt\} \qquad k > 0 \tag{A.2}$$

Firm 2 has a t_0 time lag, so its costs are:

$$g' = \frac{m}{r+k}\dot{Y}\exp\{-kt\}\exp\{kt_0\} \tag{A.3}$$

The production path of firm 1 satisfies equation (8) which is:[10]

$$X = \frac{m}{2n}(1 - \exp\{-kt\}) \tag{A.4}$$

This means firm 1, starting from zero output at $t=0$, increases production over time, approaching asymptotically an output $m/2n$. The latter is the equilibrium level a profit-maximizing monopoly will attain in a static model.

Firm 2 finds that marginal revenue is less than marginal cost as in equation (9). This condition, together with (A.4), results in the inequality:

$$m - \frac{m}{2}(1 - \exp\{-kt\}) \leq m\exp\{-kt\}\exp\{kt_0\} \tag{A.5}$$

So long as t fulfills the above inequality for given k and t_0, firm 1 will be a natural monopolist.

III. *Imposed Monopoly*

A weaker form of monopoly is an im-

[9] Mathematically, a declining marginal cost due to scale can be represented by: $g_X(X, 0, t) \leq g_X(0, 0, t)$. Usually, $g_X(0, 0, t) \leq g_Y(0, 0, t-t_0)$ because of the technological lead the monopolist enjoys. Therefore, $g_X(X, 0, t) \leq g_Y(0, 0, t-t_0)$. Nevertheless, generally $g_X(X, \dot{X}, t) > g_X(X, 0, t)$ for $\dot{X}\neq 0$. But the scale economy will probably dominate under normal circumstances so that $g_X(X, \dot{X}, t) \leq g_Y(0, 0, t-t_0)$.

[10] The corresponding price path is:
$P = m - m/2\ (1 - \exp\{-kt\})$

posed monopoly which can be defined as: A firm is defined to have an imposed monopoly for a period of time when the firm is producing *not* according to its profit maximization criterion *but to keep* the output of a potential second firm at zero. In the context of the model presented, firm 1 will be an imposed monopolist if it uses a strategy such as a limited price to keep the potential competitors out of the market after t_0; that is, for $t>t_0$, firm 1 will set the price at:

$$(11)\quad \frac{f(X,0,t)}{r} \leq \frac{g_Y(0,0,t-t_0)}{r} + g_{\dot{Y}}(0,0,t-t_0) - \frac{g_{\dot{Y}t}(0,0,t-t_0)}{r};\ t>t_0$$

while the monopolist is not maximizing his profit but still makes money *in the long run*. That is:

$$(12)\quad \int_{t_0}^{\infty} \exp\{-rt\}[Xf(X,0,t) - g(X,\dot{X},t)]dt > 0$$

This type of monopolistic behavior implies that the monopolist is willing to forego some profit for security. Note that firm 1 by assumption is a protected monopolist for $t<t_0$ and thus can theoretically produce according to equation (8). However, to carry out a strategy for $t>t_0$, firm 1 may have to adjust the production schedule for $t<t_0$ also.

For $t>t_0$ firm 1 is an imposed monopolist not maximizing its profit. Therefore, in addition to (11) and (12), the following must also hold:

$$(13)\quad \frac{f(X,0,t)+Xf_X}{r} \leq \frac{g_X(X,\dot{X},t)}{r} + g_{\dot{X}} - \frac{g_{\dot{X}t}}{r} - \frac{g_{\dot{X}X}\dot{X}}{r} - \frac{g_{\dot{X}\dot{X}}\ddot{X}}{r}\quad t>t_0$$

Equation (13) means that the MR of firm 1 must for at least some time be lower than its MC. Note that this inequality sign is the key condition that differentiates a natural from an imposed monopoly.

With certain cost patterns, it is possible for an imposed monopoly to become a natural monopoly again. An example demonstrating the possible existence of this transition is presented later.

A further condition may be added. In order that the strategy not be abandoned at some time $t_1(t_1>t_0)$, the present value of the strategy at that time must also be positive, or:

$$(14)\quad \int_{t_1}^{\infty} \exp\{-rt\}[Xf(X,0,t) - g(X,\dot{X},t)]dt > 0$$

It is quite clear that this does not imply that $Xf(X,0,t)>g(X,\dot{X},t)$ for either all t or $t>t_1$.[11] On the other hand, the condition $Xf>g$(or $g/X<f$) at all times is sufficient to satisfy (14). Using this sufficient condition along with (11) yields:

$$(15)\quad \frac{g(X,\dot{X},t)}{X} < f(X,0,t) \leq g_Y(0,0,t-t_0) + rg_{\dot{Y}}(0,0,t-t_0) - g_{\dot{Y}t}(0,0,t-t_0)\quad t>t_0$$

Thus, a sufficient but not necessary condition for an imposed monopoly is that the monopolist's average cost is less than the marginal cost less the change in marginal cost for any potential competitor.

The conditions for such an imposed monopoly and the transition from an imposed to a natural monopoly are now demonstrated by an example.

Example B. Assume a linear demand function:

[11] Readers can try $Xf-g=r+\sin t$ to confirm this statement.

$$(B.1) \qquad P = m - nX - nY$$

The cost function is assumed to be only a function of changing production levels. These costs are assumed to be increasing over time as the firm is forced to use either machines more difficult to adjust or higher priced labor to adjust them. Specifically, the cost function for firm 1 is assumed to be:

$$(B.2) \quad g = a \exp' \{b\dot{X}t\}; \qquad a, b > 0$$

The potential competitor is assumed to be *not* restricted as quickly; its cost is assumed to be lower than the monopolist's cost for all periods:

$$(B.3) \qquad g' = a \exp' \{b\dot{Y}t - b\dot{Y}t_0\}$$

Until time t_0, firm 1 is a protected monopolist. A policy decision is, however, implemented to prevent entry by competitors after t_0. As will be seen, this decision required certain actions of the firm before expiration of its protected monopoly, but for the moment, we will consider the time after t_0.

After time t_0, the monopolist acts to satisfy (11) (*using the equality sign only for simplification*). This yields:

$$(B.4) \quad X = \frac{m + ab + rabt_0}{n} - \frac{rabt}{n} \quad \text{for } t > t_0$$

Note that for this example, if firm 1 has a natural monopoly, it will produce $X = m/2n$ at $t=0$ and continue at that output forever according to (8). For a firm which is not a natural monopoly, it has to produce according to (B.4) which means that $X=(m+ab)/n$ at period t_0 and $\dot{X}=-rab/n$ from then on. Given the demand curve this means a negative price, and thus losses for the monopolist.[12] As time goes on, X decreases and eventually the firm moves back into the black.

Another condition firm 1 must meet is that its MR is less than MC as specified in (13). For this example, this means:

$$(B.5) \quad m - 2nX \leq ab \exp' \{b\dot{X}t\} (rt - 1 - bXt)$$

This inequality eventually, at some time t_1, is unsatisfied. Using (B.4), the value of t_1 is found to be the root of the equation:

$$(B.6) \quad \frac{-m - 2rabt_0 - 2ab + 2rabt_1}{-ab + rabt_1 + \dfrac{ra^2b^3}{n}} = \exp' \left\{\frac{-rab^2t_1}{n}\right\}$$

After t_1, the firm will have a natural monopoly. The production path after t_1 can therefore be solved by using (8) which yields:

$$(B.7) \quad m - 2nX = ab \exp' \{b\dot{X}t\} \cdot (rt - 1 - b\dot{X}t - bt^2\ddot{X}) \qquad t > t_1$$

Of course, before t_0, when firm 1 had a protected monopoly, it also followed this equation to determine output, though the boundary conditions are $X(0)=0$ and $X(t_0)=(m+ab)/n$ and thus the production path is different from that adopted by a natural monopoly. The difference arises in the constraints applied to production levels at t_0 and t_1, as well as in the production path followed between t_0 and t_1. Without the threat of outside competition the monopolist would maximize profits forever by producing $m/2n$. With the threat, he prepares himself during the period of patent protection to drive out

[12] The reason that negative price results is that both firms are assumed, following Cournot, to have long-run marginal and average costs of zero. Note that an addition of fixed unit cost to prevent the negative price would not add anything to the analysis but rather would obfuscate the point desired; a loss taken by a monopolist may be a perfectly rational (although not necessarily optimal) decision under certain conditions.

opposition later by changing his production schedule (and thus his price) to prevent competition. Finally, he changes his schedule again because the threat is removed. So long as the present value at t_0 is positive, which depends on the parameters, this can be a perfectly rational course to follow. In fact, the result is not surprising because it appears strikingly like many a nineteenth century monopolist.

IV. *Collusion and Oligopolistic Behavior*

Instead of competing, the monopolist may collude with the potential entrant for their mutual benefit. The objective is then to maximize the combined present values rather than for each firm to maximize on its own. The division of profits between them—a problem very much emphasized by studies in oligopolistic behavior—need not concern us except as it affects decisions implemented prior to t_0.

The objective at t_0 is clear; the combined present values are to be maximized. That is;

$$\text{(16)}\quad \max_{X,Y} \int_{t_0}^{\infty} \exp\{-rt\}[(X+Y)f(X,Y,t) - g(X,\dot{X},t) - g(Y,\dot{Y},t-t_0)]dt$$

For the case where the maximization is done over two decision variables, the Euler condition becomes two relations. In this case, these are:

$$\text{(17)}\quad f + (X+Y)f_X = g_X + rg_{\dot{X}} - g_{\dot{X}X}\dot{X} - g_{\dot{X}\dot{X}}\ddot{X} - g_{\dot{X}t}$$

$$f + (X+Y)f_Y = g'_Y + rg'_{\dot{Y}} - g'_{\dot{Y}Y}\dot{Y} - g'_{\dot{Y}\dot{Y}}\ddot{Y} - g'_{\dot{Y}t}$$

Because $f_X = f_Y$, the preceding system of equations implies:

$$\text{(18)}\quad (g_X - g'_Y) + r(g_{\dot{X}} - g'_{\dot{Y}}) - (g_{\dot{X}t} - g'_{\dot{Y}t}) = (g_{\dot{X}X}\dot{X} - g'_{\dot{Y}Y}\dot{Y}) + (g_{\dot{X}\dot{X}}\ddot{X} - g'_{\dot{Y}\dot{Y}}\ddot{Y})$$

This is a differential equation condition involving only *cost* parameters and is analogous to the static two plant monopoly case. (See, e.g., [5, p. 196].)

The following example is used to illustrate this case:

Example C. Assume the demand function is again linear;

$$\text{(C.1)}\quad P = m - nX - nY$$

The cost functions are related only to changes in output and are identical:

$$\text{(C.2)}\quad g(X,\dot{X},t) = \dot{X}^2/2$$
$$g(Y,\dot{Y},t) = \dot{Y}^2/2$$

Until time t_0, firm 1 has a protected monopoly. The boundary conditions are therefore $X(0)=0$, $X(\infty)=m/2n$. The production path from 0 to t_0 is then:

$$\text{(C.3)}\quad X = \frac{m}{2n}\left(1 - \exp\left\{\frac{rt - t\sqrt{r^2+8n}}{2}\right\}\right) \qquad t \leq t_0$$

After t_0, the two firms collude and behave according to the differential equations in (17). The initial condition for $X(t_0)$ is given by (C.3) and for $Y(t_0)=0$. The end condition at $t=\infty$ is $X+Y=m/2n$. The production paths for both firms are therefore:

$$\text{(C.4)}\quad X = \frac{m}{4n}(2 - \alpha - \beta)$$
$$Y = \frac{m}{4n}(\alpha - \beta) \qquad t > t_0$$

where

$$\alpha = \exp\left\{\frac{r - \sqrt{r^2+8n}}{2}t_0\right\}$$

$$\beta = \exp\left\{\frac{\sqrt{r^2+16n} - \sqrt{r^2+8n}}{2}t_0 + \frac{r - \sqrt{r^2+16n}}{2}t\right\}$$

The ultimate outputs, asymptotically approached, are:

$$\text{(C.5)} \quad X = \frac{m}{4n}(2 - \alpha), \qquad Y = \frac{m\alpha}{4n}$$

Thus, with this optimal collusion strategy, the share of each firm will depend on the length of time firm 1 had a monopoly.

The collusion strategy followed by the monopolist may be harmful to his interests. He bargains from less strength than he might if he increased output before t_0. The results generated are however due to the cost and demand functions chosen and the collusive agreement.

A relevant question is whether the consumer will enjoy lower price than if a monopoly is continued after t_0. It is clear that, asymptotically, the consumer is exactly as well off in each case. However, for this example, the total output from the duopoly is:

$$\text{(C.6)} \quad X + Y = \frac{m}{2n}(1 - \beta) \qquad t > t_0$$

The total output, had firm 1 remained a monopoly, would be given by (C.3). Because the value of (C.6) is greater than the value of (C.3) at any time, the consumer will pay lower prices and thus be better off under the collusive duopoly than under monopoly. Therefore, even if the asymptotical result of a dynamic collusive duopoly is the same as the static equilibrium result, the production paths indicate that the consumer will enjoy a lower price.

It is also feasible to extend the dynamic duopolistic collusion concept in the same way dynamic monopoly is extended in Sections II and III. The duopoly may be a natural duopoly or an imposed duopoly. There may even be collusive triopoly solutions. Carrying the extension to the ultimate may be thought of as giving an explanation for all conditions from monopoly to perfect competition. However, we note that not only do the equations become rapidly more complex as the number of firms becomes larger, but also that stronger and more tenuous assumptions are needed. For instance, regional and other differences in cost, not allowed in the model, are likely to assume a greater role. Assumptions will also have to be made on the conjectural variations of each firm. Hence, the area is open for further research.

V. *Conclusion*

This paper constructs a dynamic duopoly model. The model contains three equations: (1) a price equation applicable to both firms which specifies price as a function of time and total quantity produced by both firms and (2) a cost equation for each of the two firms specifying total cost as a function of quantity produced, the rate of change of quantity produced, time and the lead one of the firms enjoys. Present value maximization is then used as the decision criterion to derive two differential equations from which the time paths of quantity produced and price charged by each firm can be determined provided that the conjectural variations are specified.

The model is then used to determine the conditions under which a natural or an imposed monopoly can exist. It is shown that while scale economy can be a significant factor for a natural monopoly to exist, such factors as the time lead a firm has, the expected shift in cost of changing production levels, etc., can play an equally important role in the maintenance of monopoly power. It is further shown that, for an imposed monopoly to be effective, the monopolist may have to prepare himself during the period he has a protected monopoly by changing his production path and thus his price path to drive out his competitor. Examples are presented to illustrate these points.

This paper finally compares the effi-

ciency of collusive duopoly with that of a monopoly. It also discusses the different problems one has to face if the dynamic duopoly model presented herein is to be further extended to study polyopoly problems.

Finally, it appears that many kinds of dynamic monopoly and entry problems can be studied in great depth within the context of the model. For example, the question of what effect a rapid growth of demand will have on monopoly power can be studied by specifying the form of the price equation more explicitly. The effect of wage rates and raw material ownership on monopoly power can also be studied by specifying the form of the cost equation. The conditions under which both a natural and an imposed monopoly cannot be maintained to prevent entry can also be studied by specifying the forms of all equations and if necessary by introducing some public policy constraints. Indeed, many problems of dynamic nature can either be attacked within the context of the model or by a slight alteration of the functional relationships.

REFERENCES

1. J. S. Bain, *Barriers to New Competition.* Cambridge, Mass. 1956.
2. ———, *Industrial Organization.* New York 1959.
3. H. Brems, *Product Equilibrium under Monopolistic Competition*, Cambridge, Mass. 1951.
4. E. H. Chamberlin, *The Theory of Monopolistic Competition.* Cambridge, Mass. 1933.
5. K. J. Cohen and R. M. Cyert, *Theory of the Firm: Resource Allocation in a Market Economy.* Englewood Cliffs, N. J. 1965.
6. M. Connors and D. Teichroew, *Optimal Control of Dynamic Operations Research Models.* Scranton 1967.
7. D. Dewey, *Monopoly in Economics and Law.* Chicago 1959.
8. W. Fellner, *Competition Among the Few.* London, 1965.
9. C. C. Holt, F. Modigliani, J. F. Muth and H. A. Simon, *Planning Production, Inventories & Work Force.* Englewood Cliffs, N. J. 1960.
10. F. Modigliani, "New Developments on the Oligopoly Front," *Jour. Pol. Econ.*, June 1958, *66*, 215–32.
11. J. Robinson, *Economics of Imperfect Competition.* London 1946.
12. G. J. Stigler, *The Theory of Price*, 3rd ed. New York 1966.
13. R. Triffin, *Monopolistic Competition and General Equilibrium Theory.* Cambridge, Mass. 1949.

Differences in Efficiency Among Industries in Developing Countries

By Mark R. Daniels*

One persistent question which comes up in the literature on international trade concerns the Heckscher-Ohlin theorem and the discussion surrounding comparative advantage and differences in relative factor endowments. Much of the discussion stems from the Leontief paradox and the large number of papers offered in explanation of the apparent contradiction to the general equilibrium explanation of differences in comparative costs among countries.[1] Among these it has been correctly re-emphasized [6] that in reality relative prices may be influenced by conditions other than factor proportions. In particular, differences in relative efficiency among countries and industries may be one of these influences. Empirically very little is known about the extent and nature of international differences in efficiency and the purpose of this paper is to present some preliminary statistical evidence on the question.[2]

The concept of efficiency employed here is a familiar form of residual where efficiency is defined and measured as any variation in output per worker unexplained by weighted differences in the capital-labor ratio. Specifically the problem becomes one of providing estimates of the neutral scalar in the constant elasticity of substitution (CES) production function [2]. The weighting scheme in this case is a combination of parameters explicitly involving certain substitution and distribution characteristics of the production process. Since the development of the function is complete in the literature, the model will only be outlined. After a discussion of some data problems an array of efficiency estimates is presented for a group of manufacturing industries in a number of developing countries. An analysis of variance of these estimates substantiates the hypothesis that industrial efficiency appears to vary systematically by country, indicating that, in general, the majority of industries in one country are more efficient than those in another. Further testing shows the emergence of three or, perhaps, four distinct efficiency groupings within which no distinct pattern emerges but between which the hypothesis of systematic variation is borne out.

I. *The Basic Model*

For each industry we begin with a degree one homogeneous production function in two variables, capital and labor, written as:

$$(1) \qquad V = \gamma[\delta K^{-\rho} + (1 - \delta)L^{-\rho}]^{-1/\rho},$$

where V is the amount of value added in the production process. The parameters δ and ρ are constants while γ, the neutral efficiency measure in which we are in-

* The author is assistant professor of economics at Brown University, on leave as a Brookings Economic Policy Fellow for the academic year 1968–69. He is indebted to his colleagues, F. X. Colaço and M. B. Schupack for helpful comments on the form and substance of this paper. Thanks are also due for useful comments made by H. Louis Stettler III, Irma Adelman, Kelvin Lancaster, and a referee for this journal.

[1] See, for example, the discussions in Caves [4, pp. 273–81] and Clement, *et al.* [5, pp. 98–104].

[2] It should be noted that the first evidence of the kind presented here was given by K. Arrow *et al.* Their results did show the presence of significant efficiency variations over a small number of countries and industries [2, p. 235].

terested, is free to vary across countries.

Assuming the productivity of raw materials to be the same across countries (by virtue of which assumption the production function is written in terms of value added) and assuming equilibrium to prevail in competitive markets where labor's and capital's real wages equal their marginal products, it is possible, except for estimating the value of ρ, to solve (1) algebraically for the values of γ in each country. This is, in fact, the procedure used by Arrow, *et al.* [2], and later by Minhas [17].

The value of ρ is obtained from a side relation where

$$\rho = (1/\sigma) - 1. \tag{2}$$

Estimates of σ, the elasticity of substitution, are obtained for the ith industry from a log linear regression of the average product of labor on the wage rate given by the following equation:[3]

$$\log\left(\frac{V}{L}\right)_{ij} = a_1 + \sigma \log w_{ij} + \epsilon_{ij} \tag{3}$$

$j=1 \cdots m$, where m is the number of observations (countries) in the sample.

Differentiating (1) with respect to L and K and equating the partial derivatives to the ratio of factor prices, we can solve, with the aid of the homogeneity condition, for:

$$\delta = \left(\left[\frac{K}{L}\right]^{\rho} \cdot \left[\frac{V - wL}{wL}\right]\right) \cdot \left(1 + \left[\frac{K}{L}\right]^{\rho} \cdot \left[\frac{V - wL}{wL}\right]\right)^{-1}. \tag{4}$$

For each industry we have an estimated value for σ. Substituting, then, for σ in (2) we can solve for values of ρ. With a measure of capital input, labor input, value added and the total wage bill for each country, (4) can be solved for δ. Finally, values for δ and ρ along with the corresponding measures of capital, labor, and value added may be substituted in (1) in order to determine a value for γ by country in each industry.

II. *Comments on the Data*

Countries in the Analysis

The countries from which data were selected and their corresponding census years are Argentina (1954), Chile (1957), El Salvador (1956), Korea (1958), Paraguay (1955), Peru (1956), Portugal (1958), and Spain (1961) [23] [24] [25] [27] [28] [29] [30] [31]. The sample size was initially defined to include only the range of developing countries—that is, most countries except those in Northern and Western Europe, Canada, the United States, Japan, Australia, and New Zealand. (Countries in the Communist Bloc were not considered.) All of the countries chosen exhibited essentially the same structural characteristics and had a per capita income level below $500 a year. It should also be noted that by limiting the sample in such a way the problem of heterogeneity in industrial output at the two-digit level is minimized, thus adding more credibility to the subsequent notion of a production function for each of these industry groupings.[4]

Data on value added and wage rates are not too difficult to come by for a large number of countries. Unfortunately, how-

[3] It turns out this form of the regression equation is incorrect except when there are no efficiency (γ) differences among countries. This can be seen by differentiating (1) with respect to L and writing:

$$\log\frac{V}{L} = a_2 + \sigma \log (w \cdot \gamma^{\rho}), \text{ where } a_2 = -\sigma \log (1 - \delta).$$

Clearly, if the wage rate varies positively with efficiency then the correct value of the independent variable in (3) is understated in high-wage countries as it may be overstated in low-wage countries. As a consequence, if there are variations in efficiency, the estimates of σ given by (3) are overestimates of the true σ.

ever, estimates of capital inputs are far more scarce and since such observations are needed for the proposed measurement of efficiency differences, the sample was necessarily reduced to those countries for which such series are available. As a result of this restriction the sample size was cut to eight countries, not surprisingly providing too few observations for significant estimates of the regression coefficient in (3). Consequently, wherever possible, each country in the sample has been represented by data from the territories, provinces, or states within the country.

The Problem of Exchange Rates

In the work done by Minhas, data were converted to U. S. dollars by using official exchange rates and, in a case of multiple exchange rates, the free market rate. For countries on a system of multiple rates the problem with using the free rate—when available—is that it, too, is often an administered rate intended to apply to transactions such as unlicensed capital transfers not otherwise covered by the schedule of multiple rates. Since most of the countries studied by Minhas do not have multiple rate structures, the issue of using free rates did not pose a particular problem. In any case, he generally sidesteps the issue by assuming wage rates to be unrelated to deviations of the actual exchange rate from the true equilibrium.

With the exception of El Salvador and Portugal, however, all the countries studied here were under a system of multiple rates in the years for which census data are available. For these countries there is a good reason for rejecting the use of free rates. Argentina, Chile, and Paraguay are the only countries for which a free rate is actually quoted and this is an administered rate in all three cases. For the remaining countries there is little to choose between the various official rates listed and certainly no apparent way of choosing the rate that best approximates the true equilibrium rate.[4]

In lieu of using a schedule of official and free rates, it was possible to construct two alternative sets of exchange rates. The first is a series based on the dollar conversion rates for statistics on world exports,[5] and the second, the one which is used, is based on a purchasing power parity calculation.[6] The use of export conversion rates was rejected primarily because such rates appear to be, for the most part, weighted averages of the multiple rates quoted by each country. Furthermore, if we assume the parity calculation to be even a rough approximation of the equilibrium rate of exchange, it is obvious that Argentina, Chile, and Paraguay have significantly overvalued exchange rates.[7] In fact, this is supported by observing that many of the converted wage figures are much higher than we would reasonably expect in these countries.

[4] For a discussion of the exchange rate structure of the various countries in the sample, see the Country Notes in the I.M.F., *International Financial Statistics* [26] for any of the years immediately after the date for which the particular census material is available.

[5] See [32] [33]. No rates were quoted for Argentina, Chile, Paraguay, or Korea since these countries already express the value of their exports in dollars. For all but Korea the rates were determined by taking the ratio of the dollar value of exports to the value of exports expressed in local currency. In the case of Korea the rate was given by a schedule of implicit conversion rates in *International Financial Statistics*, XII, No. 10, Oct. 1959, p. 177.

[6] See [34, pp. 444–45]. The rates quoted in the above source, based on an assumed 1938 equilibrium are calculated only up to 1953. To bring the conversion rates up to date the 1953 rate was multiplied by the ratio of wholesale price indices (1953=100) between the United States and the country in question. The ratio of cost of living indices was used where no indices of wholesale prices were available.

[7] For Argentina the calculated purchasing power parity rate of exchange in U. S. cents per unit of national currency is 4.942 while the corresponding export conversion rate is 15.193. A similar comparison for Chile is 114.287 to 190.934 and for Paraguay, 1.393 to 3.323.

In spite of its limitations, most writers acknowledge that the purchasing power parity may provide a reasonable first approximation to the equilibrium rate of exchange, especially in periods characterized by marked changes in the price level.[8] The importance of choosing the appropriate exchange rate is indicated by computing the regression coefficient in (3) using data converted to U. S. dollars by, first, the purchasing power par rate and then the export conversion rate. Since the number of observations differs within countries not even the ordinal ranking of the regression coefficients is preserved and the value of the elasticity of substitution based on the export conversion rates is significantly higher in all but six of the industries considered.

A Measure of Capital Input

The problem of obtaining a tenable measure of the contribution of capital to the production process continues to be the most serious stumbling block in this kind of analysis. Once the sample is restricted to developing countries, then, even the conventional measures, gross or net book value of capital stock, are usually unavailable.

The measure of capital input used here is the rated horsepower capacity of installed prime movers and electric motors in operation.[9] The primary reason for choosing this particular proxy for capital stock is that fairly complete figures exist and are more or less independent of the accounting and collection techniques employed in the census divisions of the various countries under consideration. The reliability of these particular figures depends only upon the ability of the firms in question to record the rated horsepower which the manufacturer specifies on his equipment. That this may be a crude measure of the contribution of capital to the production process is undeniable. On the other hand, in a sample restricted to manufacturing industries rated capacity may give a reasonable indication of the capital input at least as reliable as the rough estimates of capital stock sometimes made available.

It has been argued that energy consumption provides a good index of the capital stock [7] [10] [20] [22]. Industries that are the largest consumers of fuel and power are also the most capital intensive. But even having demonstrated a tenable relationship between energy consumed and capital stock, it does not necessarily follow that there is the same relation between rated capacity and capital stock. This is due to the fact that the conversion of energy to work differs on the average between prime movers and electric motors. Since there is no reason to presume that the mix of the two classes of machinery is identical across countries, the use of rated capacity may bias the measure of capital input even if energy consumption is in a one-to-one correspondence with capital stock. Again the apologist's view is that for any given industry the prime mover-electric motor mix is roughly the same and because the appropriate energy to work conversion ratios are similar, about 25 to 30 per cent for prime movers and 35 to 40 per cent for electric motors, any substantial bias is unlikely to be introduced by using rated capacity in place of energy consumed.[10]

[8] See, for example, Haberler, [12 pp. 45–51]. More recently a comprehensive discussion of the problem appears in Holmes [13, 686–95].

[9] A prime mover is considered as a device for converting energy in nature directly into the energy of motion. Electric motors do not fall into this category insofar as they consume electric energy which itself must be first produced by a prime mover. From the standpoint of the firm, however, electric motors operated by power purchased externally may be thought of as prime movers. Further problems with measurements are discussed in the text.

[10] At the risk of being carried away by our apparent good fortune in finding a viable index of capital stock, it would be well to call attention to another of the obvious shortcomings of the measure. Clearly the contribution of nonenergy using capital, namely land and plant, is being ignored in the implicit assumption that its contri-

Some additional care must be taken with using rated capacity since there could be a strong element of double counting contained in the measure. This arises because much electric power is produced in steam-generating plants and it turns out that most aggregate horsepower figures available count not only the output of electric power but also the output of the prime movers used to produce that power. For the industrial sector as a whole this limitation can cast considerable doubt on the reliability of the rated capacity measure as an index of capital stock or mechanization since the industrial sector as a whole includes the electric power industry. For the manufacturing sector alone, however, this problem is not likely to cause too much trouble since most firms can be expected to buy their electric power from external sources; that is, from a centralized electric power industry in a particular region. To the extent firms employ any equipment for generating power internally there will be an element of double counting in the rated capacity figures. This is so since firms recording such figures include not only the horsepower of, say, a diesel operated generator but also that of any electric motors which operate on the electricity produced by that generator. The available statistics provide no way of determining to what extent this problem may or may not exist in the present circumstances. None of the census material surveyed provides information on what proportion of total electric power was purchased externally for the purpose of operating electric motors, a figure which would be a correct one and would avoid double counting.

When dealing with developing countries it is tempting to argue that since these countries are relatively capital poor they probably make complete or nearly complete utilization of their capital stock. Unfortunately this does not always seem to be the case, but in the usual absence of appropriate data no attempt has been made to develop an adjustment factor for determining the percentage of capital input being effectively utilized in the production process. The problem is compounded further by the fact that the degree of utilization differs between prime movers and electric motors. Even on the assumption that the prime mover-electric motor mix is the same by industry across countries it is still necessary ideally to find a factor for adjusting one or the other element of rated capacity in terms of relative differences in the effective degree of utilization between the two. Again no such correction proved possible and apart from noting the problem the data were left unadjusted.

III. *Estimates of Efficiency*

Table 1 presents estimates of the elasticity of substitution for 17 industrial groupings based on the U.N. two-digit International Standard Industrial Classification.[11] Utilizing the estimates in Table 1

bution to the production process is minimal and identical for each industry across countries. If this is not true, then there may be a further bias introduced into any efficiency estimates based on rated capacity. Imagine, for example, an industry with the same rated capacity and labor force in each of two countries. In one case the equipment is installed in a modern plant while in the other it is installed in an antiquated building. In all probability the former country would appear more efficient when this might not be the case at all since the effective capital-labor ratio is being substantially understated by using rated capacity as a measure of capital stock. In reality, however, such polar cases are not likely to occur, particularly at the manufacturing industry level, and consequently, it is possible to again reasonably vindicate the proxy for capital stock. At the same time, however, it must be reemphasized that what we have here as a measure of capital stock is only a statistical artifact. The question is only how good a statistical proxy do we have and not how good a theoretical notion of capital do we have? In fact, some studies of engineering production functions at the firm level indicate a trade-off between energy consumption and capital stock for a given level of output.

[11] The regression equation from (3) is given for each industry as

and based on the procedure described through equations (4), Table 2 provides estimates of the efficiency parameters for selected two-digit industries.[12] Since the purpose of this study is to provide some insights into the nature of international differences in industrial efficiency the next step is, therefore, to provide some analysis of the efficiency estimates. To this end the array of coefficients in Table 2 actually tempts a row-by-row comparison which would not be valid. An efficiency measure of .75 for Chile in Industry 20 only has meaning when compared to other countries within the same industry. Any attempt to compare the efficiency measure for Chile in Industry 20 with that of, say, Chile in Industry 21, or indeed any other industry or country outside 20, will not lead to meaningful or useful conclusions.

This point is simply another way of viewing the neutrality criterion. Within an industry, where the other parameters in the production function are assumed to be constant, one country may be viewed as relatively more efficient than another if its efficiency parameter has a higher value. Between industries at least one parameter in addition to the efficiency measure may be different, making direct comparisons of the scalar, in general, therefore, impossible.[13] Suppose, for example, there are two

$$\log\left(\frac{V}{L}\right)_{jk} = a + \sigma \log w_{jk} + e_{jk}$$

$j = 1 \cdots m$ where m is the number of countries;
$k = 1 \cdots l$ where l is the number of territories, or states in the j^{th} country.
V/L = thousands of dollars of the value added per man year,
w = average wage per man year in thousands of dollars.

[12] While the estimates of σ given in (3) are clearly biased if γ varies across countries, it turns out that the computed values of γ are relatively stable and generally insensitive to changes in σ. To test this point, efficiency parameters were calculated based on a number of values of σ ranging from one-half to twice those in Table 1. Chosen in this way the range of values of σ for each two-digit industry grouping encompasses nearly all similar cross-section estimates provided by Minhas [17], Minasian [16], Bell [3] and others summarized in Nerlove [19]. The results indicate that these substantial changes in the elasticity of substitution make relatively little difference to the efficiency estimates. In most cases the γ values differ by only a few per cent and the ordinal ranking from most to least efficient is completely preserved.

It should be noted at the same time that the insensitivity of the measures of γ to changes in σ does not come about for compelling economic reasons. Indeed, the degree of sensitivity is a function of the units of measurement of capital and labor as well as the extent to which the capital-labor ratio in any one country differs from a particular "critical" value. To illustrate this point, (1) may be rewritten as:

$$\gamma = y[\delta x^{-\rho} + (1 - \delta)]^{1/\rho}$$

$y = V/L$ and $x = K/L$. If $x = 1$, which is to say for any one country the units of K and L are chosen such that $K = L$, then γ will be unaffected by changes in σ (which result in changes in ρ). Once the units of K and L have been chosen then the "critical" value of K/L is established at unity. Within an industry we can imagine having chosen K and L such that for a particular country their ratio is one. But the K/L ratio differs across countries and the further the value of K/L differs from unity in the next country, the more sensitive will be its value of γ to a change in σ. Of course, we are not free to vary the units of measurement of K and L once they have been chosen but it so happens in this case, with K measured in single units of rated capacity and L measured in man-years, that most of the K/L ratios do not differ greatly from one. For those that do we find an increasing degree of sensitivity of γ to changes in σ.

[13] Following Arrow *et al.* [2, p. 235], a test of the neutrality assumption (δ constant) may be provided by rewriting (1) in the form

$$V = (\beta K^{-\rho} + \alpha L^{-\rho})^{-1/\rho}, \quad (5)$$

where $\beta = \delta\gamma^{-\rho}$ and $\alpha = (1-\delta)\gamma^{-\rho}$. From (5) if α is constant, or nearly so, efficiency variations would be concentrated in capital and, conversely, in labor for a constant β. From the definitions of β and α it is clear that if β and α vary proportionately (that is, β/α is a constant) then δ is a constant. It is possible to compute coefficients of variation to see which of α, β or δ is the more nearly constant. It turns out that for the industries in Experiment 1 in the text below the average coefficient of variation for δ is 13.34 per cent compared with values almost half again as large for α and β, 18.05 per cent and 19.24 per cent respectively. On these grounds we accept the relative constancy of δ as being a best choice of the three alternatives.

In support of the assumption of degree one homogeneity in the production function, for the majority of industries no significant correlation either positive or negative was present between the efficiency measures and scale of operations measured as average value added per firm. Sample sizes were too small to provide a decisive test but there appeared no substantial reason for rejecting the assumption of constant returns to scale on the basis of available evidence.

TABLE 1—ESTIMATES OF THE ELASTICITY OF SUBSTITUTION BY TWO-DIGIT INDUSTRY GROUPING

International Standard Industrial Classification	Sample Size	a	σ	Standard Error of σ	Coefficient of Determination R^2
20 Food industries	32	1.0538	.7506	.0819	.7367
21 Beverage industries	31	1.7418	1.3517	.2326	.5380
22 Tobacco	13	2.0415	1.4956	.2144	.8157
23 Textiles	37	.8752	1.0151	.0738	.8439
24 Footwear, other wearing apparel, and made-up textile goods	54	.7983	.7883	.1453	.3615
25 Wood and cork, except furniture	50	.8059	.8640	.0845	.6852
26 Furniture and fixtures	25	.5183	.8025	.1651	.5069
27 Paper and paper products	22	1.2193	1.3394	.1592	.7798
28 Printing, publishing and allied industries	48	.6760	.8209	.0781	.7059
29 Leather and leather products except footwear	43	.6315	.5334	.1019	.4004
30 Rubber products	18	1.4347	1.3074	.1304	.8627
31 Chemicals and chemical products	41	1.2383	1.0918	.1134	.7037
33 Nonmetallic mineral products	50	1.0889	1.1151	.0682	.8477
34 Basic metal industries	18	1.5451	1.8023	.3851	.5779
35 Metal products, except machinery and transport equipment	23	.7435	.9708	.1820	.5753
37 Electrical machinery and apparatus	33	.6906	.3826	.1624	.1518
38 Transport equipment	24	.2832	.4695	.1059	.4722

Source: Individual country data [23] [24] [25] [27] [28] [29] [30] [31].

industries each of which is characterized by a Cobb-Douglas production function. In both cases the elasticities of substitution are the same but the elasticities of output with respect to inputs differ; that is, imagine Chile in Industry 20 having the production function $V=.5K^{.6}L^{.4}$ and in Industry 21 having the production function $V=.7K^{.3}L^{.7}$, as illustrated in Figure 1.

In this case it is not possible to argue

TABLE 2—EFFICIENCY ESTIMATES

Industry	Argentina	Chile	El Salvador	Korea	Paraguay	Peru	Portugal	Spain
20	(*)	.7528	.8497	.4297	.3606	.3919	(*)	.6459
21	(*)	.8989	2.1883	2.1984	.6660	.9122	(*)	1.9944
22	9.1964	14.4979	2.4244	(*)	2.6473	(*)	5.4062	11.2670
23	.9132	.7874	.7258	.8707	.6703	.7239	(*)	1.5028
24	3.5582	1.5528	1.7979	2.7468	1.8669	.3760	(*)	2.5851
25	.4674	.4269	.7360	.5815	.3046	.3916	.6435	.8443
26	(*)	.9130	1.8002	1.3013	.3478	.9164	(*)	1.5270
27	.4596	.3140	.4215	.5128	.3877	.5626	.3048	.8978
28	1.4867	1.4820	2.2985	1.4259	1.8234	1.4238	(*)	2.2088
29	1.1811	.5728	.8837	1.6723	.9635	.5003	(*)	1.1829
30	.7034	.9946	.7756	.5833	.1799	.9757	.6289	1.3226
31	.7854	.7961	1.2611	.3790	.4377	.8581	(*)	1.3145
33	.5033	.6350	.7169	.6920	.3784	.5408	(*)	.9205
34	(*)	.6516	1.2629	.7082	(*)	3.1396	(*)	1.0936
35	(*)	.5959	1.4269	.6984	.5376	1.1391	(*)	(*)
37	1.3428	2.1809	7.1362	1.7630	1.3170	.6965	(*)	(*)
38	(*)	1.1718	1.7626	.7405	.3642	1.7406	(*)	1.1649

Source: Individual country data [23] [24] [25] [27] [28] [29] [30] [31]. (*) indicates data were not available.

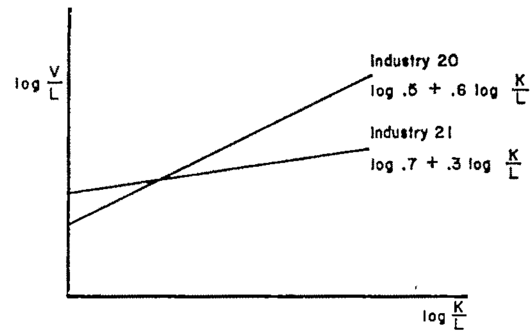


FIGURE 1. AN EXAMPLE OF NONCOMPARABILITY OF EFFICIENCY COEFFICIENTS

that Chile is more efficient in 21 than in 20 since the efficiency measures are not comparable. Figure 1 makes it obvious that whether Chile is more efficient in 21 than in 20 depends not only upon the relative height of the production function (the neutral efficiency measure) but also upon the particular value of the capital-labor ratio. We can say a country is unequivocally more or less efficient in one industry than in another only if differences in γ affect both factors identically leaving the ratio of their marginal products a constant. Even normalizing each of the rows in Table 2 in terms of one country would not improve the comparability of the estimates.[14]

Thus, the problem is to find a common unit of account to compare the efficiency coefficients in Table 2. One solution is to provide a test based on efficiency ranks. It is possible to rank the efficiency estimates within each industry starting from most to least efficient and ask whether or not there is any significant difference in the mean efficiency ranking between countries. In other words, if the estimates in Table 2 are transformed into an array of ranks, do the countries in the sample appear to be systematically more efficient or less efficient or do they all have about the same number of relatively efficient and and inefficient industries when compared to one another?

This amounts to a problem in the analysis of variance based on a two-way classification of the efficiency ranks—by industry and by country.[15] From Table 2, however, there are some missing observations and while the technique for dealing with these in a completely randomized design is easy to handle, for randomized blocks more than one or two missing observations cannot be dealt with adequately [15, pp. 229–32]. As a consequence of this difficulty, for analytical purposes Table 2 may be broken down to yield three separate and overlapping experiments, each one designed so as to contain no missing ob-

[14] This issue of noncomparability is an important one since it involves establishing the basis for a particular analytical approach to be used in providing evidence on systematic efficiency variation across countries. At the risk of laboring the issue perhaps one simple parallel example will serve to clarify any additional questions on the point. Suppose there are four or five individuals each of whose daily diet is made up entirely of, say, four different kinds of fruit. Each person has some of each fruit but they all have different amounts. A researcher wants to know if there is any significant difference in the nutritional value of the diet available to each individual. Certainly one question he can ask is whether or not there is a significant difference in the mean number of pieces of fruit in each diet. The problem with asking such a question, however, is that it provides no real information on the comparative nutritional value of the diets. Obviously what is needed is a common denominator in which to express the various elements in each diet, like calories or some appropriate measure.

Suppose now we come back to the coefficients in Table 2 and ask the question whether or not there is any systematic variation in efficiency across countries. This is analogous to the question asked by the researcher in the preceding example. The temptation is to add up the various columns of efficiency coefficients and test to see if there is a significant difference in mean efficiency units across countries. This is, however, a solution identical to the one in the example and in the absence of some common denominator such a test, like comparing mean numbers of pieces of fruit, provides no useful information on systematic efficiency variation.

[15] A test very similar to variance analysis on ranked data is available through computing Kendall's coefficient of concordance [14, pp. 80–98]. In this case, however, since we require the variance analysis framework for further tests it is convenient to adopt it throughout this section.

servations. The experiments are given as follows:

1. To contain the largest possible complete block of ranks; this covers industries 23 through 33, excluding 26 and 32, and across each of the countries except Portugal.
2. To include Portugal in as many industries as possible; namely 25, 27, and 30, Industry 22 being excluded since other observations are missing.
3. To see what effect excluding Spain would have on the conclusion about systematic efficiency variations. This is done since it seems from Table 2 as if Spain ranks as the most or near most efficient in the majority of industries. For this reason it was felt that Spain might be enough to render a significant difference in mean rank and it would be interesting to see if, excluding Spain, results would be the same. This experiment covers industries 23 through 37 excluding 26, 32, 35, and 36 and across each of the countries except Portugal and Spain.[16]

The figures in parentheses in Table 3 indicate the rank equivalents of the three experiments.

The hypothesis to be tested in all three cases is that there are no significant differences in mean efficiency ranking by country. The assumptions under which it is possible to test this hypothesis require that each of the observations (ranks) be drawn from normal populations with the same variance. To satisfy these conditions while preserving the ranking derived from the experiments selected from Table 2, a normal score transformation can be applied to the ranks. Table 3 gives the results of such a transformation. This transformation is designed for use on ranked data and through it any such population may be transformed into a normal population with mean equal to zero and variance equal to one. The normal scores are themselves the n averages of the ranked observations of all possible samples of size n drawn from a normal population with mean equal to zero and variance equal to one [15, pp. 517–22].

In testing the hypothesis that there is no significant difference in mean efficiency ranking the statistic

$$F = \frac{\text{Country mean square}}{\text{Error mean square}}$$

with $k-1$ and $(k-1)(n-1)$ degrees of freedom may be used. For each experiment k is equal to the number of countries and n is equal to the number of industries. The computing method for the analysis of variance of the normal scores is quite simple since both the grand total of all observations as well as the individual industry totals are equal to zero. The industry effects are thus completely eliminated. Table 4 provides a summary of the analysis of variance in each of the three experiments.

In all cases the F-value is significant at the 5 per cent level, which means there is a difference in mean ranks and efficiency may be viewed as varying systematically across the countries in the sample. The result also holds at the 1 per cent level but

[16] Strictly speaking it is only legitimate to drop selected rows and columns in Table 3 from the analysis if they are independent of one another. This restriction holds up all right by industry but by country some difficulties arise. Each of the efficiency estimates in an industry depends upon a certain value of σ which is determined from the regression equation (3). By excluding a particular country the regression analysis should actually be repeated without including that country's observations on V/L and w. In this case, however, we only exclude Spain in one experiment, Portugal in another, and both in a third. This means, at most, two observations in the regression analysis (since no provincial data were available in Spain or Portugal). Thus, the regression coefficient would be unlikely to change significantly and, given the insensitivity of γ to changes in σ, certainly the efficiency ranking would be preserved.

TABLE 3—NORMAL SCORE TRANSFORMATIONS OF THE RANK EQUIVALENTS OF THREE EXPERIMENTS DESIGNED TO TEST FOR SYSTEMATIC EFFICIENCY VARIATION

Experiment 1—Largest complete block:

Industry	Argentina	Chile	El Salvador	Korea	Paraguay	Peru	Spain
23	.76 (2)	.00 (4)	−.35 (5)	.35 (3)	−1.35 (7)	−.76 (6)	1.35 (1)
24	1.35 (1)	−.76 (6)	−.35 (5)	.76 (2)	.00 (4)	−1.35 (7)	.35 (3)
25	.00 (4)	−.35 (5)	.76 (2)	.35 (3)	−1.35 (7)	−.76 (6)	1.35 (1)
27	.00 (4)	−1.35 (7)	−.35 (5)	.35 (3)	−.76 (6)	.76 (2)	1.35 (1)
28	.00 (4)	−.35 (5)	1.35 (1)	−.76 (6)	.35 (3)	−1.35 (7)	.76 (2)
29	.35 (3)	−.76 (6)	−.35 (5)	1.35 (1)	.00 (4)	−1.35 (7)	.76 (2)
30	−.35 (5)	.76 (2)	.00 (4)	−.76 (6)	−1.35 (7)	.35 (3)	1.35 (1)
31	−.35 (5)	.00 (4)	.76 (2)	−1.35 (7)	−.76 (6)	.35 (3)	1.35 (1)
33	−.76 (6)	.00 (4)	.76 (2)	.35 (3)	−1.35 (7)	−.35 (5)	1.35 (1)

Experiment 2—Including Portugal:

	Argentina	Chile	El Salvador	Korea	Paraguay	Peru	Portugal	Spain
25	−.15 (5)	−.47 (6)	.85 (2)	.15 (4)	−1.42 (8)	−.85 (7)	.47 (3)	1.42 (1)
27	.15 (4)	−.85 (7)	−.15 (5)	.47 (3)	−.47 (6)	.85 (2)	−1.42 (8)	1.42 (1)
30	−.15 (5)	.85 (2)	.15 (4)	−.85 (7)	−1.42 (8)	.47 (3)	−.42 (6)	1.42 (1)

Experiment 3—Excluding Spain:

	Argentina	Chile	El Salvador	Korea	Paraguay	Peru
23	1.27 (1)	.20 (3)	−.20 (4)	.64 (2)	−1.27 (6)	.64 (5)
24	1.27 (1)	−.64 (5)	−.20 (4)	.64 (2)	.20 (3)	−1.27 (6)
25	.20 (3)	−.20 (4)	1.27 (1)	.64 (2)	−1.27 (6)	.64 (5)
27	.20 (3)	−1.27 (6)	−.20 (4)	.64 (2)	−.64 (5)	1.27 (1)
28	.20 (3)	−.20 (4)	1.27 (1)	−.64 (5)	.64 (2)	−1.27 (6)
29	.64 (2)	−.64 (5)	−.20 (4)	1.27 (1)	.20 (3)	−1.27 (6)
30	−.20 (4)	1.27 (1)	.20 (3)	−.64 (5)	−1.27 (6)	.64 (2)
31	−.20 (4)	.20 (3)	1.27 (1)	−1.27 (6)	−.64 (5)	.64 (2)
33	−.64 (5)	.20 (3)	1.27 (1)	.64 (2)	−1.27 (6)	−.20 (4)
37	−.20 (4)	.64 (2)	1.27 (1)	.20 (3)	−.64 (5)	−1.27 (6)

TABLE 4—ANALYSIS OF VARIANCE FOR THE NORMAL SCORE RANK EQUIVALENTS

Experiment	Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F	Significance Level of F for Appropriate Degrees of Freedom	
						at 5 per cent	at 1 per cent
1. Largest complete block	Industry	0	0				
	Country	19.6373	6	3.2729	6.10	3.40	3.20
	Error	25.7695	48	.5369			
	Total	45.4068	54				
2. Including Portugal	Industry	0	0				
	Country	10.7866	7	1.5409	3.03	2.77	4.28
	Error	7.1072	14	.5707			
	Total	17.8938	21				
3. Excluding Spain	Industry	0	0				
	Country	9.5804	5	1.9161	2.72	2.42	3.45
	Error	31.6696	45	.7038			
	Total	41.2500	50				

only in Experiment 1; in the other two cases the F value is not significant. Generally if we view the countries and industries in the sample as a random selection from a larger group of countries and industries, with similar structural characteristics, then the results in Table 4 may be interpreted as indicating that industrial efficiency does differ systematically across developing countries.

Having found the k country mean ranks to be significantly different at the 5 per cent level is not enough to establish that all k means are different from one another. Further tests are necessary in order to determine which of the k country means are equal and which are different. One such test was tried in Tables 3 and 4, Experiment 3, but its effectiveness is limited and a more general test is available. This is the New Multiple Range Test developed by Duncan [8]. Basically it involves testing for significant differences between all possible pairs of country means. For the data in Experiment 1 using a shorthand notation for each country consisting of its first three letters, the mean normal scores to be compared from highest to lowest are:

SPA	ELS	ARG	KOR	CHI	PER	PAR
1.11	.25	.11	.07	−.31	−.50	−.73

Table 5 gives the results of the New Multiple Range Test of the data in Experiment 1. In general, there are $\frac{1}{2}k(k-1)$ mean comparisons to be made. The g refers to the number of means (inclusive) contained in the range of any two groups (countries) to be compared. SSR refers to the shortest significant range and is obtained by multiplying the significant studentized ranges for a 5 per cent significance level with 48 degrees of freedom by the standard error of the mean (which is given as the square root of the error mean square divided by n, the number of industries in the sample).

For each country in Table 5 the mean differences are calculated as the largest minus the smallest, the largest minus the second smallest and so on. With one exception the differences are said to be significant if they exceed the critical value given by the corresponding shortest significant range. The one exception is that no comparison within a nonsignificant group can be called significant. For this reason once a nonsignificant comparison is identified,

TABLE 5—NEW MULTIPLE RANGE TEST OF THE MEAN NORMAL SCORES IN EXPERIMENT 1

g	Comparisons	Difference	SSR	Conclusion
7	SPA-PAR	1.84	.80	significant
6	SPA-PER	1.61	.78	significant
5	SPA-CHI	1.42	.77	significant
4	SPA-KOR	1.04	.76	significant
3	SPA-ARG	1.00	.73	significant
2	SPA-ELS	.86	.70	significant
6	ELS-PAR	.98	.78	significant
5	ELS-PER	.75	.77	nonsignificant
4	ELS-CHI	—	—	nonsignificant
3	ELS-KOR	—	—	nonsignificant
2	ELS-ARG	—	—	nonsignificant
5	ARG-PAR	.84	.77	significant
4	ARG-PER	—	—	nonsignificant
3	ARG-CHI	—	—	nonsignificant
2	ARG-KOR	—	—	nonsignificant
4	KOR-PAR	.81	.76	significant
3	KOR-PER	—	—	nonsignificant
2	KOR-CHI	—	—	nonsignificant
3	CHI-PAR	.42	.73	nonsignificant
2	CHI-PER	—	—	nonsignificant
2	PER-PAR	—	—	nonsignificant

no further tests need be made between any two means which are in the range of the means contained in the nonsignificant comparison. For example, the nonsignificant difference between El Salvador and Peru makes comparisons between El Salvador and Chile, Korea, or Argentina unnecessary since these are all within the limits established by the initial comparison.

The results in Table 5 may be conveniently summarized in the following schema:

SPA	ELS	ARG	KOR	CHI	PER	PAR
1.11	.25	.11	.07	−.31	−.50	−.73

Again using the abbreviated notation for each country, a line is drawn under each pair of countries exhibiting a nonsignificant difference in aggregate efficiency ranking (mean normal score). One line may contain within it several nonsignificant pairs such as, for example, the segment representing the nonsignificant difference between El Salvador and Peru containing similar comparisons between El Salvador and Argentina and Argentina and Korea among others.

The test indicates that Spain, as probably would be expected, is significantly more efficient than all the other countries in the sample. In the opposite extreme Paraguay is significantly less efficient than most of the other countries with the exception of Chile and Peru. Within the seven countries, then, there clearly emerge three distinct efficiency groupings, the extremes containing Spain on the one hand and at least Paraguay on the other. There is insufficient evidence to establish whether Chile or Peru belong to the least efficient group, the middle group containing El Salvador, Argentina and Korea, or a distinct group of their own. Only further empirical evidence can clear up this point.

REFERENCES

1. R. G. D. ALLEN, *Mathematical Analysis for Economists*. London 1960.
2. K. J. ARROW, H. B. CHENERY, B. S. MINHAS AND R. M. SOLOW, "Capital-Labor Substitution and Economic Efficiency," *Rev. Econ. Stat.*, Aug. 1961, *43*, 225–50.
3. F. W. BELL, "The Role of Capital-Labor Substitution in the Economic Adjustment of an Industry Across Regions," *So. Econ. Jour.*, Oct. 1964, *31*, 123–31.
4. R. E. CAVES, *Trade and Economic Structure*. Cambridge 1960.
5. M. O. CLEMENT, R. L. PFISTER AND K. J. ROTHWELL, *Theoretical Issues in International Economics*. Boston 1967.
6. S. CLEMHOUT, "Production Function Analysis Applied to the Leontief Scarce-Factor Paradox of International Trade," *Manchester School*, May 1963, *31*, 103–13.
7. G. H. DANIEL AND E. A. G. ROBINSON, *Planning of Investment in Fuel and Power in Underdeveloped Countries*, Section A, Paper 15, World Power Conference, Belgrade, Yugoslavia 1957, pp. 221–43.

8. D. B. DUNCAN, "Multiple Range and Multiple F Tests," *Biometrics*, 1955, *11*, 1–42.
9. R. A. FISHER, *Statistical Methods for Research Workers*, 3rd ed. rev. New York 1958.
10. A. G. FRANK, "Industrial Capital Stocks and Energy Consumption," *Econ. Jour.*, March 1959, *69*, 170–74.
11. V. R. FUCHS, "Capital-Labor Substitution: A Note," *Rev. Econ. Stat.*, Aug. 1963, *45*, 463–68.
12. G. HABERLER, *A Survey of International Trade Theory*, Special Papers in International Economics, No. 1, rev., Princeton 1961.
13. J. M. HOLMES, "The Purchasing-Power-Parity Theory: In Defense of Gustav Cassel as a Modern Theorist," *Jour. Pol. Econ.*, Oct. 1967, *75*, 686–95.
14. M. G. KENDALL, *Rank Correlation Methods*. New York 1955.
15. J. C. R. LI, *Statistical Inference I*. Ann Arbor 1964.
16. J. R. MINASIAN, "Elasticities of Substitution and Constant-Output Demand Curves for Labor," *Jour. Pol. Econ.*, June 1961, *69*, 261–70.
17. B. S. MINHAS, *An International Comparison of Factor Costs and Factor Use*. Amsterdam 1963.
18. R. R. NELSON, "The CES Production Function and Economic Growth Projections," The RAND Corporation, mimeo. P. 2942, Santa Monica 1964.
19. M. NERLOVE, "Recent Empirical Studies of the CES and Related Production Functions," in *The Theory and Empirical Analysis of Production*, NBER Stud. in Income and Wealth, no. 31, New York 1967, pp. 55–121.
20. F. R. OLIVER, "Industrial Capital Stocks and Energy Consumption—A Comment," *Econ. Jour.*, Dec. 1959, *69*, 809–12.
21. R. M. SOLOW, "Technical Change and the Aggregate Production Function," *Rev. Econ. Stat.*, Aug. 1957, *39*, 312–20.
22. J. TAYLOR, "A Surrogate for Regional Estimates of Capital Stock," *Bul. Oxford Inst. Econ. Stat.*, Aug. 1967, *29*, 289–90.
23. ARGENTINA, DIRECCION NACIONAL DE ESTADISTICA Y CENSOS, *Censo industrial 1954*. Buenos Aires 1960.
24. CHILE, MINISTERIO DE ECONOMIA, DIRECCION DE ESTADISTICA Y CENSOS, *III Censo nacional de manufacturas*. Santiago de Chile 1960.
25. EL SALVADOR, MINISTERIO DE ECONOMIA, DIRECCION GENERAL DE ESTADISTICA Y CENSOS, *Segundo censo industrial y commerical 1956*, Vol. 1. San Salvador 1959.
26. INTERNATIONAL MONETARY FUND, *International Financial Statistics*, 1959, *12*.
27. KOREA, THE KOREAN RECONSTRUCTION BANK, *Final Report—Census of Mining and Manufacturing 1958*. Seoul 1959.
28. PARAGUAY, MINISTERIO DE INDUSTRIA Y COMMERCIO, *Primer censo industrial*. Asunción 1958.
29. PERU, MINISTERIO DE FOMENTO Y OBRAS PUBLICAS, DEPARTAMENTO DE ESTADISTICA INDUSTRIAL, *Estadística industrial año 1956*. Lima 1959.
30. PORTUGAL, INSTITUTO NACIONAL DE ESTADISTICA, *Estadistica industrial 1958*. Lisbon 1959.
31. SPAIN, INSTITUTO NACIONAL DE ESTADISTICA. *Estadistica industrial 1961*. Madrid 1963.
32. United Nations, *Direction of International Trade*, ST/STST/SER.T/100, Dec. 1958, New York 1958, pp. 69–70.
33. United Nations, *Direction of International Trade*, ST/STST/SER.T/125, Jan 1962, New York 1962, p. 96.
34. United Nations, DEPARTMENT OF ECONOMIC AND SOCIAL AFFAIRS, *Patterns of Industrial Growth 1938–1958*, ST/STAT/SER.P/1. New York 1960.

COMMUNICATIONS

A NOTE ON UNCERTAINTY AND PREFERENCES IN A TEMPORAL CONTEXT

When someone asks you which of a set of uncertain prospects you prefer you should answer: "That depends upon when the outcomes will become known." The importance of this qualification was clearly pointed out by Markowitz [3, Chs. 10–11]. The reason why the temporal aspect is important is not only that it will affect your choice, but, more fundamentally, because it has to do with the question of whether or not it is possible to represent such choices by means of a utility function. Markowitz gives an example which illustrates that this may not be so.

If there are situations where choices among probability distributions cannot be represented by a utility function, then the rule for making those choices violates the consistency conditions of von Neumann and Morgenstern. This note examines the problem in the context of a simple two-period consumption model, with the purpose of finding out just what it is that is violated, and to see why such a violation is not unreasonable.

Consider an individual faced with the problem of allocating his wealth y between present consumption c_1 and future consumption c_2. To make things simple it will be assumed that the interest rate on savings (positive or negative) is zero with certainty. Thus, $c_2=y-c_1$. Total wealth could be conceived of as consisting of present income y_1 and future income y_2 with y_1 known with certainty and y_2 being a random variable. It is, however, only their sum that matters; the important point is that the precise value of y will not become known before the end of the initial period, i.e., after c_1 has been chosen. Dreze and Modigliani [2] refer to the probability distribution for y, $F(y)$, in such a case as a *temporal* uncertain prospect.

The determination of an optimal value of c_1 requires the existence of a cardinal utility function $u(c_1, c_2)$ representing preferences among consumption profiles (or among gambles involving consumption profiles). This function is assumed to be determined on the basis of the ordinary consistency conditions of von Neumann and Morgenstern. Thus, when the optimal c_1 is chosen, expected utility will be

$$U(F) = \max_{c_1} \int u(c_1, y - c_1)dF(y).$$

This expression defines indirectly the utility (or preference index) U associated with the probability distribution $F(y)$ and thus establishes a preference ordering over all such probability distributions. That is,

$$\max_{c_1} \int u(c_1, y - c_1)dF_1(y) \geqq \max_{c_1} \int u(c_1, y - c_1)dF_2(y)$$

if and only if $F_1(y)$ is preferred or indifferent to $F_2(y)$.

The question now is: can the preference ordering $U(F)$ be represented by means of a utility function $v(y)$ for wealth? That is, is it possible to write

$$\max_{c_1} \int u(c_1, y - c_1)dF(y) = \int v(y)dF(y) \ ?$$

Clearly, if choices among distributions $F(y)$ satisfy the von Neumann-Morgenstern conditions, it always would. However, as will be demonstrated, the answer is that such a representation does not generally exist. This result must mean that the von Neumann-Morgenstern conditions are violated, even though the individual acts in a perfectly rational and consistent manner.[1]

[1] The concept of the indirect (derived, tangential) utility functional was pioneered by Roy [6]. It has been applied to sequential decision problems under uncertainty by, among others, Radner [5] and Mossin [4], as well as in [2]. However, in none of these papers is the representation problem explicit. A general discussion of representation of preferences over distributions is forthcoming in a book by Karl Vind, University of Copenhagen.

To see what is involved, consider an example with the utility function

$$u(c_1, c_2) = 72 - (8 - c_1)^2 - (8 - c_2)^2 \qquad (c_1 \leqq 8,\ c_2 \leqq 8).$$

Substituting $c_2 = y - c_1$, this becomes

$$u(c_1, y - c_1) = 72 - (8 - c_1)^2 - (8 + c_1)^2 + 2(8 + c_1)y - y^2$$

so that expected utility is

$$\int u(c_1, y - c_1)dF(y) = 72 - (8 - c_1)^2 - (8 + c_1)^2 + 2(8 + c_1)E - E^2 - S^2,$$

where E and S are the expectation and standard deviation of y, respectively. The optimal value of c_1 is here given by $c_1 = E/2$, which, upon substitution, gives

$$U(F) = \max_{c_1} \int u(c_1, y - c_1)dF(y) = 72 - \tfrac{1}{2}(16 - E)^2 - S^2.$$

Thus, preferences among probability distributions are specified in terms of their means and standard deviations according to the formula above. This could be used to plot indifference curves for such prospects in the E, S-plane; these would be ellipses with center at (16, 0) and eccentricity 1/2. Clearly, for an arbitrary $F(y)$, no function $v(y)$ such that $\int v(y)dF(y) = 72 - \frac{1}{2}(16 - E)^2 - S^2$ exists. The only utility function giving expected utility in terms of E and S only (for arbitrary distributions) is the quadratic. In that case, however, the indifference curves are concentric circles with center on the E-axis (see Borch [1]).

The explanation of the result in the preceding paragraph is not far to seek. Consider a specific distribution $F_1(y)$ such that $y = 6$ or $y = 10$ with equal probability. This has mean $E_1 = 8$ and standard deviation $S_1 = 2$. Under this prospect maximum expected utility would be $U(F_1) = 36$. This probability distribution carries the same expected utility as the *certain* outcome $y = 16 - 6\sqrt{2} \approx 7.51$, or, to put it differently, the individual is indifferent between $F_1(y)$ and the degenerate distribution $F_2(y)$ which gives $y = 7.51$ with certainty.

Since $F_1(y)$ and $F_2(y)$ are ranked indifferently the individual may consider letting the choice be decided by help of a coin toss. This device may seem simple enough, but in a temporal context it must be used with care: it makes a great deal of difference whether the coin is tossed now or a year from now. If the coin is tossed *now* everything is fine: he will know which of the two distributions will apply; in either case he will, by choosing the appropriate c_1, obtain a utility level of 36, hence he should not mind letting chance decide which one will apply.

If the coin is to be tossed *after* c_1 has been chosen, however, he faces the temporal prospect $F(y) = \frac{1}{2}F_1(y) + \frac{1}{2}F_2(y)$ which has as outcomes

$$\begin{array}{c} 6 \text{ with probability } \tfrac{1}{4} \\ 16 - 6\sqrt{2} \text{ with probability } \tfrac{1}{2} \\ 10 \text{ with probability } \tfrac{1}{4} \end{array}$$

For this prospect, $E = 12 - 3\sqrt{2}$ and $S^2 = 36 - 24\sqrt{2}$ so that maximum expected utility is $U(F) = 19 + 12\sqrt{2} \approx 35.97$. Thus, the probability distribution F ranks below F_1 and F_2, i.e., $U(\frac{1}{2}F_1 + \frac{1}{2}F_2) < U(F_1)$ even though $U(F_1) = U(F_2)$. This represents an obvious violation of the independence condition (or "compounding of probabilities" condition) for an expected utility representation of $U(F)$. Such a representation requires that $U(\alpha F_1 + (1-\alpha)F_2) = \alpha U(F_1) + (1-\alpha)U(F_2)$ for any two distributions.

It is not difficult to see why this condition is not satisfied. In determining his optimal c_1 under the mixed prospect F the individual will take into account the need for hedging against the uncertain outcome of the coin toss. But a decision made on this basis is quite obviously optimal against *neither* of the original distributions F_1 and F_2. Relative to F_1 he will, in retrospect, have consumed too little; relative to F_2 too much.

The situation is illustrated in somewhat generalized form in Figure 1, where $g_i(c_1)$ is written for $\int u(c_1, y - c_1)dF_i(y)$. $g_1(c_1)$ and $g_2(c_1)$ attain the same maximum value; this value is higher than the maximum attained by the function $g = \frac{1}{2}g_1 + \frac{1}{2}g_2$.

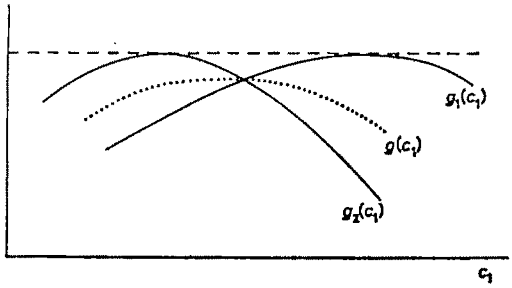


FIGURE 1

The general idea is that having to wait until the end of the period for the outcome to become known is bad enough in itself; not even knowing the exact distribution according to which the outcome will be determined is still worse. The following analogy should be familiar: I may be indifferent between teaching Course A and Course B next fall, but, in view of the preparations that must be made in the meantime, I would certainly not want the decision to be postponed until classes begin.

The example given above demonstrates the importance of the temporal aspect in risk-taking situations. As emphasized by Markowitz, the representation of preferences among probability distributions by means of a utility function is meant to apply to cases of *timeless* prospects, i.e., to situations where the outcome is determined at once, without any intervening decisions involving commitment of the outcome. For such prospects, the derivation of an indirect utility function for wealth from the solution of the allocation problem causes no difficulty (see [2]). For temporal prospects, however, representation of preferences in terms of a utility function for wealth may be inappropriate—and for reasons that are obvious once you think of it. On further reflection it is also apparent that in the real world *temporal* prospects, not timeless ones, are the rule rather than the exception. Even in a controlled experimental setting they seem difficult to avoid; in a gamble with payoffs depending upon the outcome of, say the next presidential election, subjects may easily be led to take the possibility of intermediate decisions into account. This should serve as a reminder of the constant need for caution in applying a utility function for wealth to the analysis of risk-taking behavior.

JAN MOSSIN*

* Professor of Business Administration, Norwegian School of Economics and Business Administration, Bergen, Norway.

REFERENCES

1. K. BORCH, "A Note on Uncertainty and Indifference Curves," *Rev. Econ. Stud.*, (forthcoming, 1969).
2. J. DREZE AND F. MODIGLIANI, "Epargne et consommation en avenir aléatoire," *Cahiers du Seminaire d'Econometrie*, 1966, *9*, 7–33.
3. H. MARKOWITZ, *Portfolio Selection*. New York 1959.
4. J. MOSSIN, "Optimal Multiperiod Portfolio Policies," *Jour. Bus.*, April 1968, *41*, 215–29.
5. R. RADNER, "Equilibre des marchés à terme et au comptant en cas d'incertitude," *Cahiers du Seminaire d'Econometrie*, 1966, 9, 35–52.
6. R. ROY, "La distribution du révénu entre les divers biens," *Econometrica*, July 1947, *15*, 205–25.

EXTERNAL DISECONOMIES, CORRECTIVE TAXES, AND MARKET STRUCTURE

This note is presented as a contribution to the continuing dismantling of the Pigovian tradition in applied economics, defined here as the emphasis on internalizing externalities through the imposition of corrective taxes and subsidies. My central point is much more elementary than those advanced by some of the other contributors to the recent discussion. R. H. Coase [1] demonstrated the inherently bilateral aspects of any externality relationship, and he showed that applying the Pigovian policy norms in neglect of the two-sidedness of the account may reduce rather than increase efficiency. Davis and Whinston [2] concentrated on the impossi-

bility of determining the size of a corrective tax that would lead to an efficient outcome under conditions of reciprocal externalities when production functions are nonseparably related. Plott [3] called attention to the necessity of identifying properly the aspect of the production process that generates the externality. I shall demonstrate that (1) even if the directional gains-from-trade are such that an orthodox corrective tax would increase efficiency, and (2) even if production functions are separable, and (3) even if no changes in the input mix are technically possible, the imposition of a corrective tax (under external diseconomy) will often reduce rather than increase welfare in the Pareto-efficiency sense. Only when the industry generating the external diseconomy is competitively organized can the corrective tax be unambiguously hailed as welfare-improving, even in the presence of all of the other required conditions. Under monopolistic organization, the corrective tax may well lead to a reduction in welfare rather than an increase.

My criticism is aimed more at the "Pigovian tradition" than at Pigou himself. His whole analytics, and that of Marshall, was implicitly based on the assumption of competitive structures, as, indeed, some of the contributors to the externality literature seem to have recognized.[1] It is necessary to distinguish, however, between the relevance of market structure for the emergence of externality and the relevance of market structure for the application of the Pigovian policy norms. For example, Ellis and Fellner state that "the 'atomistic' character of one producer's output under competition, frequently thought to be crucial in the external economies-diseconomies context, is not decisive of itself" [5, p. 262]. Ellis and Fellner were referring here to the potential for the emergence of externalities, but it is relatively easy to see how this statement could be taken to imply that market structure also has little relevance to the application of the standard externality-correcting devices. And we know that the levy of corrective taxes under diseconomies and the provision of corrective subsidies under economies have been widely discussed without reference to market organization. This attitude is surely characteristic of modern treatments of pollution control. If, as I shall demonstrate, it is necessary to limit the Pigovian correctives on the tax side to situations of competition much of the current discussion on these problems requires substantial revision. As we recognize, most of the problems falling under "congestion" as a general category involve external diseconomies.

My argument can be presented geometrically in the simplest of models, one in which constant cost is assumed. More complex models are not needed. An industry demand curve is shown as D in Figure 1, with the cost curve shown by $MC(AC)$. If the industry is competitively organized, equilibrium output is Q_c, and price is P_c. Let us now assume that a "bad" is suddenly discovered to be inherent in the output of this industry, an external diseconomy that is directly related to the number of units produced and not to any particulars of the input mix or to the rate of output for any other industry. This external diseconomy does affect the production functions of all firms in a second industry, also assumed to be competitively

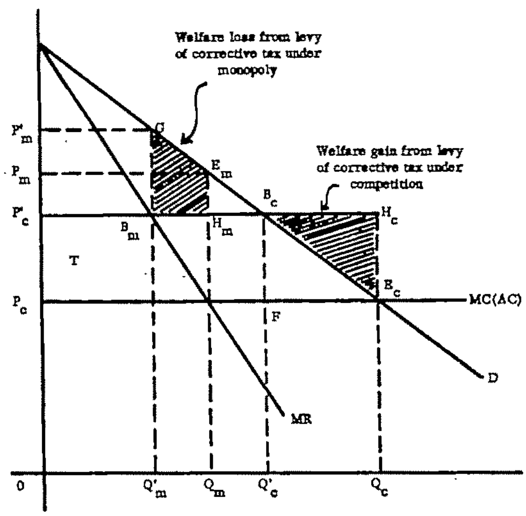


FIGURE 1

[1] Notably, Meade [4].

organized. The firms in the second industry have no legal claims to compensation for damages. Furthermore, for purposes of simplification, we assume that the costs of organizing firms in the second industry for the purpose of bribing firms in the first industry are prohibitive.

Given these restrictions, it is possible to indicate the size of a per unit tax to be imposed on the firms in the externality-generating industry. The orthodox Pigovian analysis suggests that the levy of this tax will induce behavioral changes that will move the economy to the efficiency locus. Let us suppose that the external diseconomy per unit is $P_c'P_c$, generating the unit tax T in Figure 1. Price will rise to P_c', and industry output will fall to Q_c'. How can the subsequent increase in welfare be measured? The rectangle $P_cP_c'B_cF$ represents a true "cost" that was previously treated as if it were consumers' surplus by the buyers of the first industry's product. If the proceeds of the tax are transferred to firms in the damaged industry, this now becomes consumers' surplus to the buyers of the product of this industry. If the proceeds are generally expended in the economy, these become diffused among all persons. Welfare gains and losses occur only with respect to the change in relative industry outputs. The buyers' evaluation of the quantity that was produced before the tax in the externality-generating industry but which quantity is eliminated by the tax is shown by the area under the demand curve over the range $Q_c'Q_c$, or by the area, $Q_c'B_cE_cQ_c$. The "cost" of this quantity to the community is indicated by the rectangle $Q_c'B_cH_cQ_c$. Hence, the welfare gain is shown by the shaded triangle, $B_cH_cE_c$.[3]

To this point, no problems are encountered given the restrictions initially placed on the model. However, let us now assume that the industry that generates the external diseconomy is organized as a monopoly, with a single profit-maximizing firm. Before the levy of any corrective tax, monopoly output is Q_m and price is P_m. As in the competitive case, Pigovian analysis suggests the levy of a corrective tax of T per unit of output. Monopoly output falls to Q_m' and price increases to P_m'.

It is easy to show that, under the conditions as shown in Figure 1, welfare has *decreased*, not increased as a result of the levy of the corrective tax. The cost of the change in quantity is measured as before, by the rectangle $Q_m'B_mH_mQ_m$. The evaluation of the quantity is measured as before, by the area under the demand curve, or by $Q_m'GE_mQ_m$. Since the latter area clearly exceeds the former, welfare has been reduced as indicated by the shaded area. The geometry makes clear that, in this simple case, this result must hold so long as the corrective tax, which we assume to have been estimated properly, is less than the difference between price and marginal revenue at the initial monopoly output.

As I have indicated, the point is a very elementary one. It is a particularly clear example of the theory of second-best. The monopolist simultaneously imposes two external diseconomies, at least in a general sense. He "pollutes" and hence increases costs of firms in the damaged industry. Also, however, he holds down output and hence increases costs of his product to buyers. So long as the second diseconomy is more highly valued than the former, any levy of a per unit tax on the monopolist's output will decrease total welfare. There are gains-from-trade here in two opposing directions, and there is no means of determining a priori which set of "trades" is potentially the more efficient. Conceptually, and ignoring costs of organizing, the firms in the damaged industry could bribe the monopolist to reduce output and thereby to reduce "pollution." At the same time, again ignoring the costs of organizing, the buyers of the monopolist's product could bribe the monopolist to in-

[3] If the damaged industry is identical in size to the industry that is generating the externality, and if demand and cost relationships are similar in the two industries, the welfare gain to the community can also be represented by the appropriate "welfare triangle" in a diagram depicting the situation of the other industry. The danger to be guarded against is double-counting of the same welfare gain in this procedure.

crease output. In some costless three-way negotiation process, the ultimate outcome under conditions such as those depicted in Figure 1 is the corrected equilibrium output at Q_o'.

As the construction as well as the discussion indicates, there is an important asymmetry between external diseconomies and external economies with respect to the possible offsetting welfare effects of market structure. With external economies, the provision of corrective subsidies reinforces the directional change in output that reforms in market structure would indicate to be desirable. In this case, buyers of the monopolist's own product could join forces with firms in an externally benefitted industry to bribe the monopolist to increase output.

As Coase has correctly emphasized, the whole approach of the Pigovian tradition is responsible for many confusions in applied economics that are slowly coming to be clarified. This approach involves an undue concentration on the decision-calculus of the firm or individual that is observed to be generating the external effects. Even if we disregard all problems of measurement, making the marginal private cost as faced by the decision-taking unit equal to marginal social cost does not provide the Aladdin's Lamp for the applied welfare theorist, and the sooner he recognizes this the better.

JAMES M. BUCHANAN*

* The author is professor of economics at the University of California, Los Angeles.

REFERENCES

1. R. H. COASE, "The Problem of Social Cost," *Jour. Law and Econ.*, Oct. 1960, *3*, 1–44.
2. O. A. DAVIS AND ANDREW WHINSTON, "Externalities, Welfare, and the Theory of Games," *Jour. Pol. Econ.*, June 1962, *70*, 241–62.
3. CHARLES PLOTT, "Externalites and Corrective Taxes," *Economica*, Feb., 1966, *33*, 84–87.
4. J. E. MEADE, "External Economies and Diseconomies in a Competitive Situation," *Econ. Jour.*, March 1952, *62*, 54–67.
5. H. S. ELLIS AND WILLIAM FELLNER, "External Economies and Diseconomies," *Am. Econ. Rev.*, Sept. 1943, *33*, 493–511, reprinted in G. J. Stigler and K. E. Boulding, ed., *Readings in Price Theory*, Homewood, Ill. 1952, pp. 242–63.

A MODIFIED GOLDEN RULE: THE CASE WITH ENDOGENOUS LABOR SUPPLY

I. *Introduction*

The important question of the optimal savings strategy for growth has occupied the minds of economists for some time and has been discussed repeatedly. Of course, any such study is based on a highly aggregated view of an economy with some facets of the real world emphasized and others ignored, but despite the long list of economists who have considered economically determined population growth, this feature has for the most part been ignored. This note is an attempt to rectify this omission by studying optimality when the growth rate of population is allowed to respond to economic factors.

In the standard neoclassical growth model the savings ratio does not affect the equilibrium rate of growth of output (which is equal to the exogenously given growth rate of labor) but rather determines the level of per capita income. Hence, an increase in the savings ratio has two effects on per capita consumption: it both reduces current consumption and increases future output which in turn allows an increase in future consumption. As is well known [4], a savings ratio equal to the share of output going to capital balances these effects and thus defines the highest sustainable level of per capita consumption among all balanced growth paths. But, if population growth responds to economic variables, in particular to per capita income, a change in the savings ratio has a third effect: per capita consumption is altered through the induced change in the

level of population. With this complication the Golden Rule prescription for savings is no longer optimal.

II. *Population Growth*

If population growth can be explained economically (which is a reasonable assumption in the long run[1]—and after all, where else but in the long run is it appropriate to talk of balanced growth paths? [5]), one must look to demographic work to find the form of the relationship. Obviously, population growth results from the interaction of birth and death rates so there are really two separate relationships to be considered.[2] I will follow John Buttrick's analysis [1, pp. 76–77] that, while death rates vary inversely with per capita income, the effect on birth rates is somewhat ambiguous.

I will assume that the economic effect on birth rates is of a smaller order so that overall, the relationship between population growth rates and per capita income is positive and monotonic (the effect on birth rates being only large enough to alter the second derivative—never enough to shift the first derivative of the population function).

Figure 1 illustrates the above theory. The relationship between income per capita and population growth may be divided into three stages: children are producers' goods and consumer durables (I), children cease to be producers' goods while costs of child rearing rise (II), children are durable luxury goods (III).

[1] It is certainly true that today underdeveloped countries have immediate access to medical techniques and knowledge which have evolved gradually over centuries and thus particular diseases can be eradicated almost overnight to drastically lower mortality rates. However, without an accompanying improvement in the standard of living there are limits to what can be done and, in fact, "economic conditions in the underdeveloped countries may become the chief determiner of their death rate in the near future" [6, p. 375].

[2] If labor mobility is considered, migration would be another influence which would strengthen the positive relationship between per capita income and the growth rate of population. Certainly no voluntary major migration has ever been isolated when the flow was from a high to a low income region.

III. *The Model*

A single homogeneous output Y is produced by two homogeneous factors, capital K and labor L, subject to a constant returns to scale production function.

$$Y(t) = F(K(t), L(t)) \tag{1}$$

Hence,

$$y \equiv \frac{Y}{L} = f\left(\frac{K}{L}\right) \equiv f(k)$$

where

$$f(k) \equiv F\left(\frac{K}{L}, 1\right)$$

with $f'(k) > 0$ and $f''(k) < 0$ for all positive k.

In addition I will assume that:

$$\lim_{k\to\infty} \frac{f(k)}{k} = 0 \quad \text{and} \quad \lim_{k\to 0} \frac{f(k)}{k} = \infty$$

and will define $c(t)$ to be the rate of current consumption per unit of labor input and $s(t)$ to be the current total (private and public) savings ratio. Hence, $c(t) = (1 - s(t))y(t)$ and since all savings are invested:

$$\frac{\dot{K}}{L} = s(t)y(t) \tag{2}$$

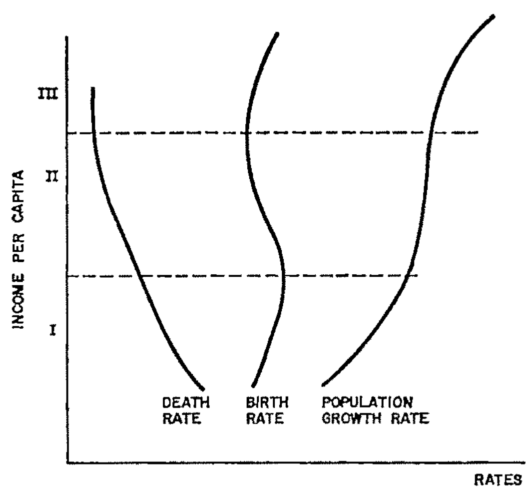


FIGURE 1. POPULATION GROWTH

The population grows at a relative rate n but here n is a function of per capita income (and it is assumed that the labor supply, which is all productively employed, grows at the same rate).

$$L = n(y)L(t) \tag{3}$$

where it is assumed that $n' > 0$.

By the assumptions made on the production function, output per unit of capital is a monotonically increasing function of the labor-capital ratio, going from 0 to ∞ as the labor-capital ratio goes from 0 to ∞. Thus, for any constant savings ratio, sy/k also goes from 0 to ∞ as $1/k$ goes from 0 to ∞. The population function is a monotonically decreasing function of the labor-capital ratio by the homogeneity of first degree assumption. Hence, there is at most one intersection between n and sy/k (an intersection which defines a balanced growth path) and there will be exactly one so long as population growth is positive for some labor-capital ratio.

Since k increases to the right of the intersection and decreases to the left of it

$$\left(\frac{\dot{k}}{k} = \frac{sy}{k} - n\right),$$

this unique balanced growth path is also a stable one. (See Figure 2.)

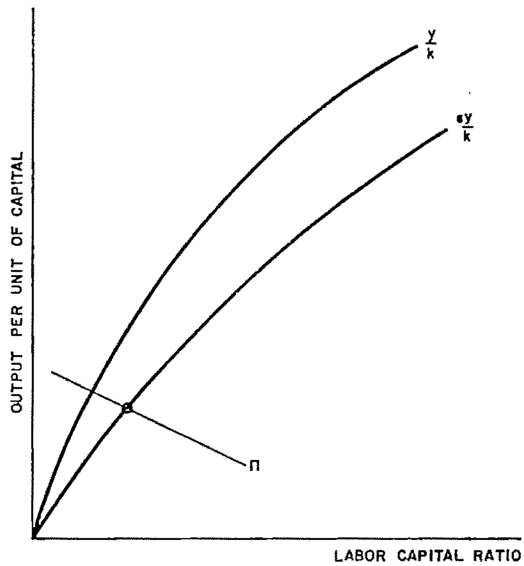


FIGURE 2. THE EXISTENCE OF A STABLE UNIQUE BALANCED GROWTH EQUILIBRIUM

IV. *A New Golden Rule of Accumulation*

Consider the problem where a constant savings ratio is to be chosen (as a policy variable) to produce maximum consumption per capita[3] among all balanced growth paths. No longer is it best to invest capital's share. Instead the optimal savings ratio is lower:[4]

$$s = \frac{kf'}{f}(1 - kn') < \frac{kf'}{f} \tag{4}$$

where kf'/f is the share of output which is paid to capital.

The second order conditions to insure that per capita consumption is in fact maximized require that the following inequality be satisfied:

$$[f''(1 - kn') - 2n'f' - k(f')^2n''] < 0. \tag{5}$$

Obviously, the existence and stability of a

[3] A discussion of the various approaches towards defining what in fact optimal growth is, can be found in [3, pp. 225–228].

[4] A balanced growth path is defined by

$$T = sf - kn = 0$$

The problem is to choose s so that $(1-s)f(k)$ is maximized subject to $sf(k) = kn(y)$ or equivalently to maximize $f(k) - kn(f(k))$ by choice of k.

$$\frac{d}{dk}[f(k) - kn(f(k))] = 0 = f'(1 - kn') - n$$

Thus, the modified Golden Rule savings ratio becomes:

$$s = \frac{kn}{f} = \frac{kf'}{f}(1 - kn') \tag{4}$$

The second order conditions for a maximum require:

$$\frac{d^2}{dk^2}[f(k) - kn(f(k))] < 0$$

and hence,

$$[f''(1 - kn') - 2n'f' - k(f')^2n''] < 0. \tag{5}$$

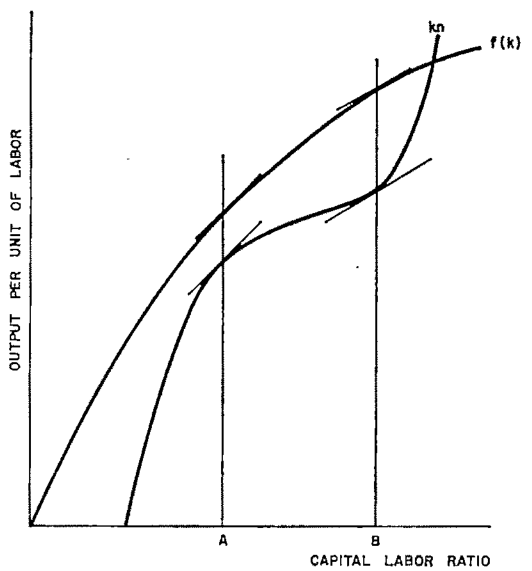


FIGURE 3. THE EXISTENCE AND STABILITY OF THE MODIFIED GOLDEN RULE

solution to this consumption maximization problem require conditions on the population function beyond the simple monotonicity which guarantees the existence and stability of a unique balanced growth solution. Since a unique intersection between kn and f is assured it is sufficient for the existence of a solution (a point where $n=f'(1-kn')$) that:

$$\frac{d(kn)}{dk} < f'(k) \tag{6}$$

where the derivatives are evaluated at that k such that $kn=0$.

There remains the possibility of multiple solutions, alternatively (local) minimum and (local) maximum. However, as long as condition (6) is satisfied there exists at least one position which is a (local) maximum. Since in Figure 3, the vertical distance between $f(k)$ and kn represents per capita consumption it is seen that A represents a (local) minimum while B represents a (local) maximum. In fact, the second order conditions define a lower bound on the second derivative of the population function, for (5) is clearly satisfied for n'' greater than or equal to zero.

These results, of course, depend crucially on the variable which is felt to influence the growth rate of population. If the growth rate of population responded not to income per capita but rather to consumption per capita, once again the optimal strategy is to save capital's share,[5] although if the population function is not monotonic, the problem of a multiplicity of solutions remains. Since the argument of the population function is the maximand, there is no change from the exogenous case—it is as if population grew exogenously at that rate corresponding to the maximum level of consumption per capita.

In summary, the introduction of endogenous population growth results in an optimal savings ratio which is less than that specified by Phelps' Golden Rule of Accumulation. Correspondingly, the growth rate of the economy is less than the marginal productivity of capital. Any optimism

[5] This result has been noted by Michio Morishima in his comments on T. C. Koopmans' paper, printed in the same volume [3, p. 295]. This conclusion may be derived as follows:

$$M = (1-s)f(k) + \lambda[sf(k) - kn((1-s)f(k))]$$

where λ is a Lagrangian multiplier.

$$\frac{\partial M}{\partial s} = 0 = -f(k) + \lambda[f(k) + kf(k)n']$$

$$\frac{\partial M}{\partial \lambda} = 0 = sf(k) - kn$$

$$\frac{\partial M}{\partial k} = 0 = (1-s)f'(k) + \lambda[sf'(k) - n - k(1-s)f'(k)n']$$

Hence, $s = kf'(k)/f(k) =$ the share of output which is paid to capital. It is sufficient for a maximum that:

$$\begin{vmatrix} M_{11} & M_{12} & T_1 \\ M_{21} & M_{22} & T_2 \\ T_1 & T_2 & 0 \end{vmatrix} > 0$$

where the subscript $\left[\begin{smallmatrix}1\\2\end{smallmatrix}\right]$ denotes partial differentiation by $\left[\begin{smallmatrix}s\\k\end{smallmatrix}\right]$. This becomes:

$$f'' < 0$$

which is exactly the same as in the exogenous population growth case. Although the existence of a Golden Rule solution with population growing, may be imperilled when population changes in response to per capita consumption, so long as the relationship is monotonic, a maximum, if it exists, is unique and there is no possibility of a solution to the first order conditions being a minimum.

generated by this lower optimal savings ratio must be tempered by bearing in mind the implicit assumption that all population (or at least a constant fraction) can and will be productively employed. To provide a basis for more meaningful policy one would certainly want to distinguish labor force from population in a more flexible way and hence allow for the possibility of "optimal unemployment" [2].

ERIC DAVIS*

* The author is an assistant professor at Purdue University. He would like to thank Professor Ryuzo Sato for his help and encouragement.

REFERENCES

1. J. BUTTRICK, "A Note on Growth Theory," *Econ. Develop. and Cult. Change*, Oct. 1960, *9*, 75–82.
2. A. R. DOBELL AND Y. C. HO, "An Optimal Unemployment Rate," *Quart. Jour. Econ.* Nov. 1967, *81*, 675–83.
3. T. C. KOOPMANS, "On the Concept of Optimal Economic Growth," in *The Econometric Approach to Development Planning*, Pontificia Academia Scientiarum, Amsterdam 1965, pp. 225–87.
4. E. PHELPS, "The Golden Rule of Accumulation: A Fable for Growth Men," *Am. Econ. Rev.*, Sept. 1961, *51*, 638–43.
5. R. SATO, "Fiscal Policy in a Neo-Classical Growth Model: An Analysis of Time Required for Equilibrating Adjustment," *Rev. Econ. Stud.*, Feb. 1963, *30*, 16–23.
6. W. S. THOMPSON AND D. T. LEWIS, *Population Problems*, New York 1965.

EXCESS BURDEN: THE CORNER CASE

The so-called "excess burden" or welfare cost of various types of taxation has been analyzed under a variety of assumptions in the public finance literature of the past thirty years. As a result of this long discussion it is now well recognized that the traditional argument for general taxes as against partial excise taxes is subject to a number of theoretical qualifications.[1] For practical purposes, however, it seems to be quite widely accepted that the traditional analysis is still highly relevant.[2] It may therefore be of interest to clarify what appears to be something of a paradox in this analysis.

A central finding in the traditional analysis of excise taxes is that the excess burden of an excise varies directly with the degree of substitutability between the taxed good and other goods. In the extreme case where there is no substitutability whatever, an excise tax has no distorting substitution effect, and excess burden is zero. When some substitutability exists, the pattern of consumption is distorted and excess burden results.

The conclusion that the magnitude of the excess burden varies directly with the degree of substitutability seems, however, to conflict with the apparent fact that the consumer suffers less of a burden in switching his consumption from the taxed good the better the substitutes that are available. In the limit, a tax on, say, one brand of toothpaste would appear to impose no excess burden at all; the consumer could if necessary switch his entire toothpaste budget from one brand to another at no welfare cost.

The paradox is, however, easily explained when we pay due attention to the corner cases which have been somewhat neglected in the excess burden literature.

Substitutability and the Corner Case

The problem can be readily illustrated without any essential loss of generality in the simple case of constant opportunity costs[3] and a linear demand schedule for the taxed product. In Figure 1, compensated demand and supply schedules for the good X are initially assumed to be D_0D_0' and S_0S_0'. In accordance with the traditional analysis,

[1] The major contributions to the lively debate over the period 1939–54 are well summarized by Walker [2].

[2] See, for example, the more recent discussions by Harberger [3, pp. 25–70], and Musgrave and Richman [3, pp. 81–131].

[3] When opportunity costs are increasing or decreasing, further issues of some complexity are raised regarding the concept of excess burden, which, however, do not affect directly the point made in the present note. See Bishop [1, Sec. IV].

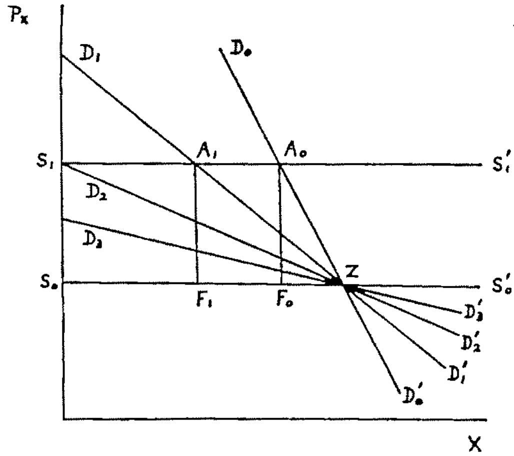


FIGURE 1

it is assumed that perfect competition prevails in all markets, factor supplies are completely elastic, there are no external economies or diseconomies and no other excise taxes in existence. A given volume of public expenditure is assumed to be financed by nondistorting general taxes. The pre-excise equilibrium point Z therefore represents a Pareto-optimal output of X.

When a per unit excise tax of amount T is imposed, the supply schedule shifts to S_1S_1', a constant distance T above S_0S_0'. The new equilibrium is at A_0. By a familiar argument, the excess burden of the excise ΔW is here given by the area of the triangle ZA_0F_0, which shows the loss of surplus suffered by consumers of X.[4] Thus

$$\begin{aligned} \Delta W &= \tfrac{1}{2}A_0F_0 \cdot F_0Z \\ &= \tfrac{1}{2}T\Delta X \\ &= \tfrac{1}{2}T^2S \end{aligned}$$

where S is the numerical value of the reciprocal of the slope of the demand schedule, i.e., F_0Z/A_0F_0, and measures the substitutability of X for other commodities.

Where substitutability between X and other commodities is zero, the demand schedule becomes a vertical straight line, $S=0$ and excess burden is therefore zero. A higher degree of substitutability is reflected in a demand schedule of slope numerically less than that of D_0D_0', e.g., D_1D_1'. As a result, excess burden rises to ZA_1F_1.

It is easy to see, however, that, as substitutability increases further and the slope of the demand schedule steadily diminishes, excess burden approaches a maximum which is reached when the demand schedule (D_2D_2') intercepts the price axis at S_1. In this latter case, the unit tax T is just sufficient to choke off consumption of X. Excess burden is then ZS_1S_0.

For substitutability greater than that represented by D_2D_2', it can be seen that excess burden *diminishes*. One such schedule is D_3D_3'. The tax T is now more than sufficient to choke off the demand for X, which cannot, however, fall below zero. As in the case of D_2D_2', all consumer surplus on X is lost, but as better substitutes are available there is now less surplus to lose; the consumer can switch his consumption away from X at a smaller welfare cost. In the limit, where perfect substitutes are available, consumption of X can be entirely abandoned without welfare cost. The general proposition that excess burden varies directly with substitutability therefore requires qualification when account is taken of corner cases.

It is also interesting to observe that *maximum* potential excess burden varies inversely with substitutability, though the achievement of these higher maxima requires successively higher rates of unit tax. Once the maximum excess burden for a given commodity has been achieved by imposing a rate of tax just sufficient to choke off consumption of that commodity, further increases in the tax rate have no effect on welfare cost.

Conclusions

When full account is taken of the neglected corner cases, we therefore see that as the degree of substitutability increases from zero through to infinity, the excess burden of an excise tax first rises from zero towards a maximum for some finite value of the substitution coefficient, and then falls back

[4] See, for example, Harberger [3, pp. 33–35].

towards zero as substitutability becomes perfect.

The intuition that the consumer should suffer less of a burden when there are good substitutes available for the taxed product seems therefore to have at least some limited analytical significance. Where the combination of the tax rate and substitutability conditions is such that consumption is reduced to zero, increases in substitutability reduce rather than increase excess burden.

For the community as a whole, such cases would, however, seem to be of only limited practical interest. Even sumptuary excises seldom aim at the complete elimination of consumption (and hence of revenue!). For practical examples we would have to look to the closely related area of fines and penalties, e.g., for the consumption of narcotics.

For the single individual the analysis may perhaps be of greater significance. Even quite modest rates of excise on a given commodity will probably reduce the consumption of certain individuals to zero. For those individuals it is then clearly the case that a higher degree of substitutability would reduce excess burden.

J. G. Head and C. S. Shoup*

* The authors are professors of economics at Dalhousie University and Columbia University, respectively.

REFERENCES

1. R. L. Bishop, "The Effects of Specific and Ad Valorem Taxes," *Quart. Jour. Econ.*, May 1968, *82*, 198–218.
2. D. Walker, "The Direct-Indirect Tax Problem: Fifteen Years of Controversy," *Pub. Fin.*, No. 2, 1955, *10*, 153–77.
3. *The Role of Direct and Indirect Taxes in the Federal Revenue System.* Princeton 1964.

SOME NEW RESULTS IN THE MEASUREMENT OF CAPACITY UTILIZATION

In a recent paper Lawrence Klein and R. S. Preston [2] provide an important review of the possibilities for measuring capacity utilization as well as a challenging assessment of the interpretations to be placed on that concept. While the analysis rests on binding assumptions such as the same rate of utilization for both capital and labor it is not at all easy to see what prospects there are for alternative functions, such as the C.E.S., giving more satisfactory results.

The main purpose of this note is to question the generality of the approach used in attempting to measure the concept of capacity utilization. The apparent failure to use the installed productive capacity may not stem solely from a lack of demand. If time is needed to learn to operate a plant there will be a running-in period before the productive capacity appears to be fully used. The usual methods of measuring capital stock and the labor force do not allow for this possibility. Certainly estimates of capital stock vary with investment outlays as well as the calculations of asset lives.

Where the growth rate of the manufacturing sector of each industry is low the effect of this learning period on estimates of capacity utilization may be small. Nevertheless, if the growth rate is high, a sizeable proportion of the capital stock may still be in this learning period. Hence an apparent underutilization may be recorded. A simple example may illustrate the point. Let m be the average life of capital, r the cumulative rate of growth of gross investment, I the annual gross investment, K the gross capital stock, x the learning period after completion of the investment outlay and before the plant is in full production so that $x<m$, and K' is the gross capital stock for only those investments in full production. Then one can postulate

$$(1)\qquad I = e^{rt}$$

$$(2)\qquad K = \int_{t-m}^{t} I dt = \frac{e^{rt}(1 - e^{-rm})}{r}$$

$$(3)\qquad K' = \int_{t-m}^{t-x} I dt = \frac{e^{rt}(e^{-rx} - e^{-rm})}{r}$$

TABLE 1—RATIO OF USED TO INSTALLED CAPITAL

		Percentage Growth Rate of Gross Investment					
		1	3	5	10	15	20
$x=1/2$	$m=10$	.9476	.9425	.9373	.9228	.9083	.8899
	$m=20$	.9725	.9670	.9610	.9436	.9250	.9068
	$m=30$	.9808	.9749	.9682	.9487	.9279	.9083
$x=1$	$m=10$	.8954	.8860	.8761	.8495	.8207	.7904
	$m=20$	.9451	.9345	.9228	.8942	.8534	.8153
	$m=30$	.9616	.9502	.9372	.9037	.8591	.8183
$x=2$	$m=10$	.7917	.7753	.7582	.7132	.6664	.6187
	$m=20$	.8908	.8709	.8495	.7904	.7272	.6642
	$m=30$	.9236	.9019	.8775	.8092	.7379	.6695

$$\frac{K'}{K} = \frac{e^{-rx} - e^{-rm}}{1 - e^{-rm}} \tag{4}$$

On these assumptions and with $r>0$ then $e^{-rx}<1$. The ratio of capital in production to the installed capital given in equation (4) is less than unity. The ratio will diminish with increases in the rate of growth of investment, r, and the learning period, x. It will rise the longer the life of plant. A series of values for this ratio is shown in Table 1 assuming different sets of values for growth rates, asset life, and the learning period. Where this period is short with values for x between $\frac{1}{2}$ and 1 the ratio may be well below unity. Hence this simplified hypothesis offers possibilities for explaining some part of apparent excess capacity.

The learning period need not be long in high income industrial countries though in some branches of manufacturing heavily dependent upon technologically oriented processes this need not necessarily apply. But some criticisms of the performance recorded by manufacturing industries in less developed countries may be misplaced if they associate underutilization of productive capacity with shortages of factor inputs [1]. In many of these countries the manufacturing sector has grown very fast so that the estimates on the right-hand side of Table 1 —10 percent growth rates or more—could be relevant to their situation. Moreover, the lack of skills in production and maintenance may shorten the length of life of plant compared with advanced industrial countries. As the data in Table 1 show, this contributes further to what appears to be the underutilization of productive capacity.

The proposition advanced in this note does little to undermine the conclusions advanced by Lawrence Klein and R. S. Preston. Indeed their scepticism about the validity of estimates showing utilization rates lower than their own series is supported for other reasons. If their high utilization periods are linked to fast investment growth the new plants will not immediately be able to achieve their long-run expected capacity. Furthermore the assumption linking the rate of unemployment directly to the rate of capacity utilization is not made to appear more rigorous, at least for industry groups.

W. P. HOGAN*

* University of Sydney, Australia.

REFERENCES

1. W. P. HOGAN, "Capacity Creation and Utilisation in Pakistan Manufacturing Industry," Economic Development Report N. 84; Development Advisory Service, Center for International Affairs, Harvard University, Cambridge 1967.
2. L. R. KLEIN AND R. S. PRESTON, "Some New Results in the Measurement of Capacity Utilization," *Am. Econ. Rev.*, March 1967, *57*, 34–58.

URBAN POVERTY AND LABOR FORCE PARTICIPATION: COMMENT

Joseph Mooney's recent article in this journal [9] reports an inverse relation between unemployment rates in urban areas and the labor force participation rates (LFPR) of white and non-white poor people in these areas. This finding augments a considerable body of other research which has found procyclical movements of LFPR's in the general population, particularly on the part of "secondary" workers. The central and novel conclusion of the study is that this negative response of labor force participation to unemployment is stronger for the poor than for the non-poor, and for non-white married women compared to white married women.

We believe this conclusion is incorrect, and in view of its significance for the positive analysis of labor supply functions and also for questions of policy, we should like to explain the source of Mooney's misinterpretation and to point out evidence that conflicts with his conclusion. The policy questions particularly concern how the labor force behavior of workers responds to changes in aggregate demand as these changes occur over time. We examine how the cross-section results Mooney reports relate to time series behavior both because of the implications for policy and because the time series provide us with a means of obtaining improved estimates of the parameters of the underlying labor supply equations.

There are two major questions that need to be asked about the meaning of Mooney's empirical findings: Can one infer from the sizes of his inter-area regression coefficients: (1) that a given change over time in aggregate demand in the economy will bring about a larger change, in the same direction, of the poor (or non-white) than of the non-poor (or white) labor force? (2) that the substitution parameter, which produces the discouragement effect of unemployment, dominates the income parameter, which produces the additional worker effect, to a greater extent in the labor supply function of the poor (non-whites) compared to the non-poor (whites)?

Our comments in the next section can be stated briefly: Contrary to the impressions conveyed by the article,[1] the answers to both questions are negative. The reasons for the negative answers can be briefly summarzied:

1. The cross-section regression estimates shown in the study are not translatable directly into aggregate time series behavior. The basic problem is that labor supply parameters enter into the empirical estimates only indirectly and in a different way in the cross-section from in the time series. Thus a wide range of behavior in time series can be shown to be consistent with Mooney's cross-section results. Indeed, direct time series observations appear to show that the cyclical labor force response of non-white married women, husband present, is weaker than that of white women, and, at times, even countercyclical.

2. The time series findings are not inconsistent with an assumption that functions of the two populations are the same. But if they are different, the findings suggest that they differ in the opposite direction from that suggested in the article, namely: the domination of the substitution parameter over the income parameter is not more, but less pronounced among the poor than among the non-poor.

3. Other cross-section studies provide both direct and inferential evidence of the difference in labor supply functions of the two populations. Most direct is the finding that for white married women compared to non-white married women (a) the negative net effect of unemployment is larger for whites, and (b) the substitution parameter is

[1] Regarding (1), see [9, last para., pp. 104, 117]. On question (2), the most explicit passage is: "It seems clear that the low income worker is more likely to be a discouraged worker than an additional worker, especially if 'she' is a non-white married woman with husband present" [9, p. 115].

As explained below, the extent to which a substitution parameter exceeds the income parameter for a worker expresses the sense in which he is more likely to be a "discouraged worker" than an "additional worker."

larger relative to the income parameter for whites regarding nontransitory earnings and income variables.

I. *Theoretical Considerations and Time Series Evidence*

As Mooney notes in his article, the effect of unemployment on the labor force participation rate (LFPR) is a mixture of two forces pulling in opposite directions: (1) the negative or "discouraged worker" effect of the demand for labor, and (2) the positive or "added worker" effect from declines in family income and the attempts of family members to find jobs to make up the loss. Given the level of wages offered (a proxy for the demand for labor) and the level of incomes received by the households, the unemployment rate can represent the transitory changes occurring in the demand for labor and in household income. It is important to keep in mind that LFPR's across areas are related to both wage and income variables, and to both normal and transitory components of these variables. The closer the model comes to specifying these variables, the more confidence we have in it.

The most important part of the specification of the model concerns the transitory or short-run changes, and we can simplify matters by first concentrating on these. The structural equation is:

$$(1)\quad Y = \alpha_1 X_1 + \alpha_2 X_2 + v \qquad \alpha_1 < 0, \alpha_2 > 0$$

where

Y = labor force participation rate
X_1 = income ("added worker") variable
X_2 = substitution ("job opportunity," or "discouraged worker") variable

The unemployment rate, U, (say, the rate for the civilian labor force) can act as a proxy variable for X_1 and for X_2, because it is correlated with both as follows (suppressing the constant term):

$$(2)\qquad X_1 = a_1 U + e_1 \qquad a_1 < 0$$

$$(3)\qquad X_2 = a_2 U + e_2 \qquad a_2 < 0$$

The empirical regression of Y on U yields a coefficient b_{YU}:

$$(4)\qquad b_{YU} = \frac{dY}{dU} = \alpha_1 \frac{dX_1}{dU} + \alpha_2 \frac{dX_2}{dU} = a_1\alpha_1 + a_2\alpha_2$$

Mooney finds that when labor force participation rates of wives (Y), in families with husband present, are regressed on the overall SMSA unemployment rate (U), in the SMSA cross-section, $b_{YU}<0$ for each of the groups white or non-white, poor or non-poor, the latter defined by average income in the urban tract. He also finds, and this is the major issue in the article and in our comment, that the negative response of the non-white (or poor) labor force to the overall SMSA unemployment rate is larger than the response of the white (or non-poor) labor force.

In symbols:

$$(5)\qquad \left| b^{n}_{YU} \right| > \left| b^{w}_{YU} \right| ,$$

where n denotes non-white (poor) labor and w denotes white (non-poor) labor. Given assumptions (1) to (4), this is equivalent to:

$$(5a)\qquad \left| a_{2n}\alpha_{2n} \right| - \left| a_{1n}\alpha_{1n} \right| > \left| a_{2w}\alpha_{2w} \right| - \left| a_{1w}\alpha_{1w} \right|$$

The point of our comment is that ***this observation in the cross-section carries no implications*** to the effect that: (a) the cyclical sensitivity of non-white (poor) labor force is greater than that of the white (non-poor) labor force, or (b) the non-white (poor) labor supply function contains a stronger substitution (or weaker income) parameter than the white (non-poor) labor supply function.

The reasons why inequality (5) need not hold in time series are indicated below. That it ***in fact does not hold in time series*** is shown in Table 1 and Chart 1. We proceed to elucidate both propositions.

Assume that supply functions for each labor force are the same, so that $\alpha_{1n}=\alpha_{1w}$, $\alpha_{2n}=\alpha_{2w}$. Under these conditions, the cross-section inequality (5a) observed by Mooney will hold if:

$$(6)\qquad \left| a_{2n} \right| > \left| a_{2w} \right| \quad \text{and}$$

$$(7)\qquad \frac{a_{2n}}{a_{1n}} > \frac{a_{2w}}{a_{1w}}$$

(See the Appendix for a demonstration of this proposition and for further comments about inequalities (6) and (7)).

Inequality (6) may hold because the general level of economic activity in an area as represented by U is likely to be a more important indicator of job opportunities for non-white than for white women. The reason is the close link between local levels of economic activity and the local demands for service and household workers, and the importance of these occupations for non-white female employment.[2]

Inequality (7) may hold because $|a_{1w}|$ is not expected to be sufficiently smaller than $|a_{1n}|$ to offset (6). Indeed, it may be that $|a_{1n}|<|a_{1w}|$, because white workers are a much larger proportion of the total labor force than non-white workers. Since the SMSA unemployment rate does not mainly, or even largely,[3] represent a cyclical position common to all groups, it is likely to be a better proxy—in a correlation and regression sense—for white than for non-white family income.

Inequalities (6) and (7) are less likely to hold in time series. Unemployment rates of whites and of non-whites move in unison over the business cycle, though with an amplitude more than twice as large for non-whites.[4] However, since average incomes (and wages) of non-whites are about one-half that of whites, the a-coefficients which are measured in dollars, may well be similar in both groups, though this can be more confidently asserted about a_1 than about a_2.

If so, and if the supply parameters are the same for the two labor forces, we should find in the time series that $b^n_{YU}=b^w_{YU}$, or the equivalent $b_{(Y_n-Y_w)U}=0$, that is to say, a similar cyclical behavior of the two labor forces. Chart 1 and Table 1B indicate, however, that the procyclical fluctuations in the labor force rates of white married women, if anything, exceed those of non-white married women. This is contrary to Mooney's findings in the cross-section. This reversal of the relation in time series could arise either because inequalities (6) and (7) are reversed in time series, and/or because the labor supply parameters (α) are different in the two groups, with the income parameters (α_1) absolutely larger (or the substitution parameter α_2 absolutely smaller) in the behavior of the non-white labor force.

The hypothesis that the income effect, α_1 ("added worker behavior"), is stronger in the poor labor force is theoretically plausible: If the "additional worker" phenomenon is viewed as an alternative to dissaving, asset decumulation, or increasing debt in the attempts of families to maintain consumption in the face of declining income, such behavior is most likely to appear in families at low levels of wealth.[5]

The differential cyclical sensitivity of Y_n compared to Y_w is shown in Figure 1 and Table 1B, based on data in Table 1A. The use of (Y_n-Y_w) is convenient: it replaces two scatter diagrams by one, and it dampens trends (which are, incidentally, similar for non-white and white married women, husband present) which would dominate the separate regressions. A fuller effect of the trend is taken into account by the regressions on lines 2 and 3 of Table 1B where time (1948=1) is included in arithmetic and logarithmic form. The partial coefficient, $b_{(Y_n-Y_w)U.T}$, remains positive and significant. A somewhat less clear but positive correlation is obtained in a similar scatter for all women, rather than married women.

[2] About 58 per cent of non-white females who live in poverty areas of the large SMSA's are service workers or private household workers. Outside these areas, 44 per cent of non-white females are in these occupations, and among white females the percentages in these occupations are 23 and 16 in poverty and nonpoverty areas, respectively. source: [11, p. 1107].

[3] On this, see Mincer [8, p. 80].

[4] The absolute and relative variances in unemployment rates are larger for non-white workers. The standard deviations of unemployment rates for non-white males and females for 1948–1966 are 2.6 and 2.0, respectively, compared to standard deviations of 1.1 and 1.0 for white males and females. The corresponding coefficients of variations are .29 and .23 for non-white males and females, respectively, and .27 and .19 for white males and females respectively. (Source of data: [5, p. 214]). For evidence on income variability among low skill groups see [10, p. 34] and [5].

[5] For an elaboration and other supporting empirical evidence see Mincer [7, p. 75] and [8, p. 95 and note 37, p. 111]. Also Cain [3, pp. 288–90].

TABLE 1A—LABOR FORCE RATES OF MARRIED WOMEN, HUSBAND PRESENT AND THE AGGREGATE UNEMPLOYMENT RATE, UNITED STATES 1948–1966

Month April (A) or March (M) and *Year*	Labor Force Participation Rates: Non-white Wives (1)	White Wives (2)	Difference (3) (1)–(2)	Aggregate Unemployment Rate (March) (4)
1948 A	30.6	21.3	9.3	4.0
1949 A	33.0	21.7	11.3	5.0
1950 M	37.0	22.8	14.2	6.3
1951 A	36.0	24.3	11.7	3.4
1952 A	33.6	24.5	9.1	2.9
1953 A	33.5	25.0	8.5	2.6
1954 A	36.9	26.0	10.9	5.7
1955 A	38.2	27.0	11.2	4.6
1956 M	40.6	28.0	12.6	4.2
1957 M	40.2	28.7	11.5	3.7
1958 M	42.0	29.2	12.8	6.7
1959 M	41.7	30.0	11.7	5.6
1960 M	40.8	29.6	11.2	5.4
1961 M	45.0	31.6	13.4	6.9
1962 M	46.5	31.5	15.0	5.6
1963 M	44.8	32.7	12.1	5.7
1964 M	45.4	33.4	12.0	5.4
1965 M	46.7	33.6	13.1	4.7
1966 M	47.6	34.3	13.3	3.8
1967 M	47.8	35.8	12.0	3.9

Source: Labor Force Rates: 1948–1958, *Current Population Reports*, p. 50; 1959–1967, *Special Labor Force Reports*.
Unemployment Rates: *Employment and Earnings*, February 1967.

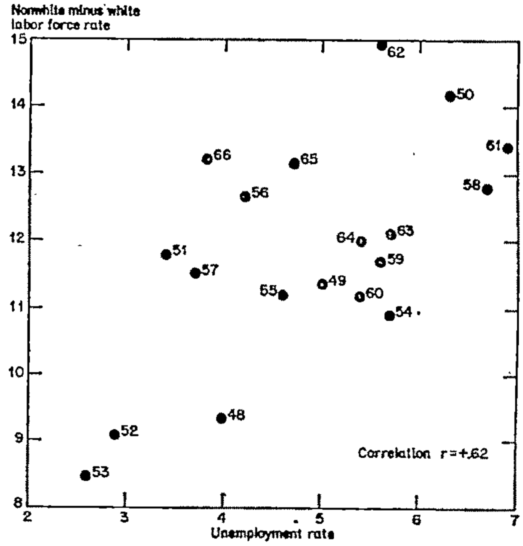


FIGURE 1.—SCATTER DIAGRAM OF UNEMPLOYMENT RATES AND THE DIFFERENCE IN LFPR'S OF MARRIED WOMEN, HUSBAND PRESENT, NON-WHITE MINUS WHITE

II. *Conflicting Evidence from Cross-Sectional Studies*

An implication of the preceding discussion is that a "cleaner" empirical specification of variables representing X_1 and X_2 will show a larger net discouragement effect for whites (or the non-poor) than non-whites (or the poor), even in cross-sections. Specifically, we have argued that the total unemployment

TABLE 1B—REGRESSION RESULTS OF THE DIFFERENCE IN LFPR'S (COL. (3) IN TABLE 1A) AND UNEMPLOYMENT RATES (COL. (4) IN TABLE 1A), WITH AND WITHOUT A TREND VARIABLE (YEAR=T)

Dependent Variable (LFPR of Non-white Wives=Yn and of White Wives=Yw)	Independent Variable (*t*-ratios in parentheses): Intercept	U	T(1948=1)	Log T	R^2
$Yn-Yw$	7.93 (6.43)	.81 (3.27)			.37
$Yn-Yw$	7.33 (6.37)	.71 (3.07)	.11 (2.20)		.51
$Yn-Yw$	6.95 (5.54)	.71 (3.00)		.70 (1.95)	.49

Source of data: Table 1A.

rate which Mooney used overstates the estimate of the "discouragement effect" and understates the estimate of the "added worker effect" for non-whites relative to whites. Indeed, the net discouragement effects that Mooney measured with the total unemployment rate are surely biased estimates for both whites and non-whites, since they are much larger than we have any reason to expect. For example, an increase in unemployment rates from 3 per cent to 7 per cent implies, by Mooney's Table 2, a decrease of 18 percentage points in the LFPR of non-white wives!

Two improvements in the specification of the statistical model are available. One is to use unemployment rates specific to the two color groups for the separate regressions with each group. The goal is to capture the income changes occurring within each color group of families (to measure α_1) and the changes occurring in the labor market for each color group of wives who work (to measure α_2). Limitations of the data, however, force us to measure the combined effects of these changes $(\alpha_1+\alpha_2)$.[6] A second improvement is to add to the model "long-run" income and wage variables and other "permanent" effects, so that the measures of short-run or transitory effects can be more nearly isolated. This step was taken in the cross-section study discussed below which shows results that differ from Mooney's. A final point which will be discussed briefly is the effect of migration on the particular cross-section regressions computed by Mooney.

[6] Our attempt with cross-section data to use simultaneously the male unemployment rate as a proxy for X_1 and the female unemployment rate as a proxy for X_2 led to ambiguous results. The "added worker effect" *was* more positive (actually, a smaller negative coefficient) for white than non-white wives. However, the "discouragement effect" was larger in absolute value for non-whites. Our opinion is that the simultaneous measure of the two effects is not feasible. Even if the unemployment rates were perfect measures of male income and female wage changes (which they are not), they are not precisely specific to the husband and wife groups studied. Furthermore, there is high collinearity between the unemployment rates, with rates for primary workers (like husbands) tending to dominate the representation of the overall state of the labor market regarding "job opportunities."

1. A direct conflict with Mooney's results is found in the study by Cain [3]. The LFPR of married women with husband present in SMSA's was found to be significantly negatively related to unemployment for white wives for both 1950 and 1960, but this relation was not significant for non-white wives. The smaller net negative effect of unemployment shown for non-white wives in this research appears correct (as shown below in Table 2), but Cain was incorrect in emphasizing for non-whites the nonsignificance of his original measure of the unemployment effect for the 1960 data. His use of the male unemployment rate introduced a bias in the opposite direction to that made in Mooney's analysis—as should be clear in our discussion above, since the "added worker effect" is exaggerated relative to the "discouragement effect" when only the male unemployment rate is used, and it turns out that this distortion is more severe for non-whites than for whites. The color-specific total (male plus female) unemployment rate appears to be the best candidate to capture the net effect of unemployment, i.e. $(\alpha_1+\alpha_2)$.

We argued above in part I that the measured (negative) effect of unemployment on the LFPR of non-white wives is exaggerated (because of the behavior of the a_i parameters) when the total unemployment rate instead of the unemployment rate for the non-white population is used. This argument is supported by cross-section regressions computed with the same data that underlie Table 2 which show that for LFPR's of non-white wives, husband present, the coefficient was -2.30 when the total unemployment rate was the regressand (regression not shown in Table 2) and -1.44 when the non-white unemployment rate was used. The coefficient for white wives was -1.79 when the white unemployment rate was the regressand. The smaller (negative) coefficient for non-white wives found when non-white unemployment rates were used is more compatible with the time series evidence.

The coefficients discussed above are all simple regression coefficients. It is with

TABLE 2.—COMPARISON OF SIMPLE AND PARTIAL REGRESSION COEFFICIENTS OF SELECTED UNEMPLOYMENT RATES; DEPENDENT VARIABLE IS THE LFPR OF WIVES, HUSBAND PRESENT, FOR NON-WHITE WIVES AND ALL WIVES, SMSA'S IN 1960

1. *Non-white wives* (All 66 SMSA's with population >250,000, with data for non-whites). LFPR of non-white wives is the dependent variable	Regression Coefficient and t-Ratios in Parentheses		$\bar{U}$	σ_u	$\sigma_u/\bar{U}$
	Arithmetic	Logarithmic[a]			
Unemployment, U, as independent variable					
(a) Total unemployment rate,					
non-whites	−1.44	−.33	9.09	2.95	.32
Simple regression	(5.65)	(5.50)			
Multiple regression[b]	−.70	−.12			
	(2.11)	(1.56)			
(b) Male unemployment rate,					
non-whites	−1.26	−.29	9.42	3.21	.34
Simple regression	(5.30)	(4.96)			
Multiple regression[b]	−.55*	−.09*			
	(1.94)	(1.26)			
(c) Female unemployment rate,					
non-whites	−1.43	−.32	8.60	2.86	.33
Simple regression	(5.45)	(5.45)			
Multiple regression[b]	−.71	−.14*			
	(1.96)	(1.77)			
2. *All wives* (same 66 SMSA's as in 1) LFPR of all wives is the dependent variable					
(a) Total unemployment rate, whites[c]	−1.79	−.25	4.26	1.15	.27
Simple regression	(5.16)	(5.28)			
Multiple regression[b]	−1.70	−.22			
	(5.19)	(4.69)			
(b) Male unemployment rates, whites[c]	−1.89	−.25	4.15	1.16	.28
Simple regression	(5.57)	(5.53)			
Multiple regression[b]	−1.79	−.24			
	(5.89)	(3.93)			
(c) Female unemployment rate,					
whites[c]	−1.21	−.18	4.51	1.35	.30
Simple regression	(3.77)	(3.99)			
Multiple regression[b]	−1.04	−.14			
	(3.25)	(3.05)			
3. *All wives* (all 100 SMSA's over 250,000 LFPR of all wives is the dependent variable)					
(a) Total unemployment rate					
(both races)	−1.41	−.26	5.14	1.54	.30
Simple regression	(6.90)	(6.95)			

* Coefficient is *not* significant at the 5 per cent level.

[a] All variables were transformed to logs except for the dummy variable for region.

[b] The other variables, besides unemployment, in the multiple regression were, for each color group, income of males, income of females, the years of schooling completed of females, the fertility rate, and a dummy variable representing the South region.

[c] Unemployment rates for whites were derived by subtracting the appropriate labor force and unemployment figures for non-whites in each SMSA from the corresponding figures for the total population.

TABLE 2 (*Continued*)

	Regression Coefficient and t-Ratios in Parentheses		$\bar{U}$	σ_u	$\sigma_u/\bar{U}$
	Arithmetic	Logarithmic[a]			
Multiple regression[b]	−1.24 (5.76)	− .22 (5.84)			
(b) Male unemployment rate			5.10	1.76	.35
Simple regression	−1.24 (6.98)	− .25 (7.36)			
Multiple regression[b]	−1.12 (5.85)	− .21 (6.25)			
(c) Female unemployment rate			5.24	1.46	.28
Simple regression	−1.11 (4.65)	− .19 (4.62)			
Multiple regression[b]	− .90 (3.98)	− .15 (4.02)			

multiple regressions, however, that we find the most important differences in Cain's results and conclusions compared with Mooney's. Table 2 shows the relation between LFPR's of wives, husband present, for the two color groups and three color-specific unemployment rates—for males, females, and the total.[7] As shown in Table 2, the simple regressions of the LFPR of non-white wives on non-white unemployment rates produce coefficients that are negative and highly significant. When the other variables are included, however, the unemployment coefficients diminish and become in some cases statistically insignificant (by slight margins).

Also shown in Table 2 are the changes between the simple and partial coefficients of the same unemployment rates (for whites) on the LFPR of total (or white) wives in the 1960 study. The differences between the simple and partial coefficients are much smaller than they were for non-whites and the partial coefficients are consistently larger for white wives than for non-white wives. Evidently the biases from the omitted variables are much stronger in the regressions of non-white wives' LFPR's. Moreover, exactly this pattern of differential effects of omitted variables is shown in Mooney's Tables 2 and 3. The coefficient of unemployment is reduced when a demand variable is added in the regressions of non-white wives, but adding the demand variable in regressions of white wives increases the coefficient of unemployment.[8]

The regression models in Table 2 included variables representing the income of the husband and the wage earnings available to wives. As reported in [3, pp. 288–90], the income parameter was larger relative to the wage parameter for non-white wives compared to white wives. These "long-run" effects corroborate the color differences found in the effects of unemployment.

The results for non-white and white wives, husband present, are suggestive of the

[7] Most of the coefficients in Table 2 for the total group of wives are different from those found in corresponding regressions that were reported in Cain [2], [3]. Table 2 benefits from a correction of a misrecorded unemployment rate for whites in one SMSA. The only qualitative correction that needs to be made in the earlier studies is to resolve the puzzle over the insignificant coefficient of unemployment in the regressions in arithmetic form—we now see that the arithmetic as well as the logarithmic form of unemployment yields significant (negative) coefficients with respect to white wives' LFPR's.

[8] It should be pointed out that Mooney was constrained from using multiple regression models by his data source, since the Census tract data do not include the same wealth of variables that are available for the entire city or SMSA. However, when Mooney did add an additional variable (one that reflected normal demand conditions for female workers in the SMSA), the gross negative effect on unemployment for non-white wives was cut by 20 per cent See his Tables 2 and 3 in [9].

results we might obtain for poor and non-poor families, but the translation is not altogether straightforward. A research focus on families in which both spouses are present provides only limited insights into many labor supply aspects of poverty issues.[9]

2. Mooney's use of poverty areas as the data base for comparing differential labor force responses to unemployment between whites and non-whites is subject to additional problems which make it risky to generalize his results to a wider or different context. The first problem, recognized by Mooney, is that poor whites are more likely to have various labor market handicaps which would explain both their presence in poverty areas and their lesser responsiveness to labor market conditions.

Another factor is a difference between white and non-white families in their patterns of residential mobility within SMSA's.[10] If better job opportunities induce and enable white families to move out of poverty areas more easily than non-white families, as is generally believed, the white population remaining in the poverty areas will show a weaker labor force response across SMSA's.

III. *Conclusion*

Mooney's research has added to our knowledge of the responsiveness of the labor force, particularly secondary workers, to unemployment conditions. The net effect of unemployment for segments of the poverty population and for non-whites is probably negative, but the effects are unlikely to be larger than those for the non-poor and white groups. Mooney's conclusions to the contrary are derived from a statistical model that is not well specified for the hypothesis he was testing, and other empirical evidence based on better specified models do not support him.

GLEN G. CAIN AND JACOB MINCER*

* The authors are respectively associate professor of economics, University of Wisconsin, and professor of economics, Columbia University and National Bureau of Economic Research. We wish to thank John Conlisk for supplying us with the proof we use in the Appendix and Albert Rees for his comments on an earlier draft. We are grateful for the clerical services and computer time made available by the Institute for Research on Poverty and by the research project, Dynamics of the Labor Market (supported by a grant from the Ford Foundation) and the Social Systems Research Institute at the University of Wisconsin.

APPENDIX

Part 1

The purpose of this part of the Appendix is to demonstrate that inequalities (6) and (7) imply that (5) or (5a) hold even if $\alpha_{1n}=\alpha_{1w}$ and $\alpha_{2n}=\alpha_{2w}$.

Given: a_{1n}, a_{2n}, a_{1w}, a_{2w}, α_1, and $-\alpha_2<0$

By assuming that the labor supply parameters between whites and non-whites are the same, we can drop the subscripts n and w on the α_i. We know:

$$\text{(A.1)} \quad a_{1n}\alpha_1 + a_{2n}\alpha_2 < 0 \quad \text{or} \quad \frac{a_{1n}}{a_{2n}} + \frac{\alpha_2}{\alpha_1} < 0$$

$$\text{(A.2)} \quad a_{1w}\alpha_1 + a_{2w}\alpha_2 < 0 \quad \text{or} \quad \frac{a_{1w}}{a_{2w}} + \frac{\alpha_2}{\alpha_1} < 0$$

If the following inequalities hold (as we argued in our comment and defend in more detail below):

$$\text{(A.3)} \quad |a_{2n}| > |a_{2w}| \quad \text{or} \quad \frac{a_{2n}}{a_{2w}} > 1$$

$$\text{(A.4)} \quad \frac{a_{2n}}{a_{1n}} > \frac{a_{2w}}{a_{1w}}$$

then:

$$\text{(A.5)} \quad (\alpha_1 a_{1n}+\alpha_2 a_{2n}) < (\alpha_1 a_{1w}+\alpha_2 a_{2w}) < 0$$

which is equivalent to inequalities (5) and (5a) in our comment. *Proof.* Inverting both sides of (A.4) and then reversing the inequality gives:

[9] Mooney also reported on the relation between LFPR's of males and unemployment. He found that the LFPR's of *males in poverty tracts* were significantly negatively affected by unemployment, but that the LFPR's of *all males* were not. Conflicting evidence is, however, reported by Bowen and Finegan [1], who report a significant negative effect of unemployment on the LFPR's of several age groups of males among 100 SMSA's in 1960. Their results are derived from a multiple regression model similar to that used by Cain.

[10] Inter-SMSA mobility, which Mooney discussed, need not be greater for poor whites than for poor non-whites, though it may be.

$$\text{(A.6)} \qquad \frac{a_{1n}}{a_{2n}} < \frac{a_{1w}}{a_{2w}}$$

Adding α_2/α_1 to both sides:

$$\text{(A.7)} \qquad \frac{a_{1n}}{a_{2n}} + \frac{\alpha_2}{\alpha_1} < \frac{a_{1w}}{a_{2w}} + \frac{\alpha_2}{\alpha_1}$$

Both sides of (A.7) are negative, from (A.1) and (A.2), so we may multiply the left side by a number>1 without violating the inequality. Let a_{2n}/a_{2w}, which is>1 by (A.3), be this number. Then:

$$\text{(A.8)} \qquad \left(\frac{a_{2n}}{a_{2w}}\right)\left(\frac{a_{1n}}{a_{2n}} + \frac{\alpha_2}{\alpha_1}\right) < \frac{a_{1w}}{a_{2w}} + \frac{\alpha_2}{\alpha_1}$$

Multiplying both sides by $\alpha_1 a_{2w}$, which is a positive number and thus does not alter the inequality, gives, after simplifying:

$$\text{(A.9)} \quad \alpha_1 a_{1n} + \alpha_2 a_{2n} < \alpha_1 a_{1w} + \alpha_2 a_{2w} < 0$$

which was to be shown.

Part 2

It may be helpful to indicate more formally the basis for inequalities (6) and (7) which are also listed above as (A.3) and (A.4). First:

$$a_{2n} = \rho_{2n}\left(\frac{\sigma_{x_{2n}}}{\sigma_U}\right) \quad \text{and} \quad a_{2w} = \rho_{2w}\left(\frac{\sigma_{x_{2w}}}{\sigma_U}\right)$$

where ρ_2 is the correlation between the transitory change in wives' earnings, X_2, and the unemployment rate, U, for the entire SMSA. We assume a bivariate normal distribution. We argued that $|\rho_{2n}|>|\rho_{1w}|$ for reasons given earlier. Moreover the variance in earnings among non-whites, $\sigma_{x_{2n}}^2$ is surely larger than that of whites. The absolute levels of earnings for white females is somewhat larger than for non-white females, but the variability of the latter group's earnings is considerably greater. Thus $|a_{2n}|>|a_{2w}|$. Second:

$$a_{1n} = \rho_{1n}\left(\frac{\sigma_{x_{1n}}}{\sigma_U}\right) \quad \text{and} \quad a_{1w} = \rho_{1w}\left(\frac{\sigma_{x_{1w}}}{\sigma_U}\right)$$

As we argued earlier, $|\rho_{1n}| < |\rho_{1w}|$, probably by a wide margin. Also, the absolute levels of non-white male incomes are no more than 60 per cent of the income levels of white males, so even though the relative variation in non-white male incomes is greater than for whites, the standard deviation of absolute income changes will be similar. That is, $\sigma_{x_{1n}} \approx \sigma_{x_{1w}}$, which implies that $|a_{1n}|<|a_{1w}|$. Even if $\sigma_{x_{1n}}>\sigma_{x_{1w}}$ and $|a_{1n}|>|a_{1w}|$, the margin would be sufficiently small so that the inequality (A.9) is preserved.

REFERENCES

1. W. G. Bowen and T. A. Finegan, "Labor Force Participation and Unemployment," in A. M. Ross, ed., *Employment Policy and the Labor Market*, Berkeley 1965, pp. 115–61.
2. G. Cain, *Married Women in the Labor Force*. Chicago 1966.
3. ———, "Unemployment and the Labor Force Participation of Secondary Workers," *Indus. Lab. Rel. Rev.*, Jan. 1967, *20*, 275–97.
4. Current Population Reports, *Income of Families and Unrelated Individuals in the United States*, Ser. P-60, U. S. Bureau of the Census.
5. L. R. Klein and N. Liviatan, "The Significances of Income Variability on Savings Behavior," *Bull. Oxford Univ. Inst. Stat.*, May 1957, *19*, 151–60.
6. *Manpower Report of the President, 1967*, U. S. Department of Labor, Washington, 1968.
7. J. Mincer, "Labor Force Participation of Married Women," in *Aspects of Labor Economics*, National Bureau of Economic Research, Princeton, 1962, pp. 63–105.
8. ———, "Labor Force Participation and Unemployment: A Review of Recent Evidence," in R. A. Gordon and M. S. Gordon, eds., *Prosperity and Unemployment*, New York 1966, pp. 73–112.
9. J. D. Mooney, "Urban Poverty and Labor Force Participation," *Am. Econ. Rev.*, March 1967, *57*, pp. 104–119.
10. J. N. Morgan, "Anatomy of Income Change," in *Consumer Behavior of Individual Families Over Two and Three Years*, (R. F. Kosobud and J. N. Morgan, eds.,) Monogr. No. 36, Survey Re-

search Center, The University of Michigan, Ann Arbor 1964.
11. "Poverty Areas of Our Major Cities," Special Labor Force Report No. 75, from the *Monthly Labor Review*, Oct. 1966.

URBAN POVERTY AND LABOR FORCE PARTICIPATION: REPLY

Several years ago, Jacob Mincer wrote a truly pioneering and classic study of the labor force participation of married women [4]. This article stimulated and fostered a great deal of useful research in the area of labor supply functions. Furthermore, Glen Cain has written the best and most comprehensive study of the labor force participation of married women [2]. Consequently, when I received this comment on my article I anticipated an illuminating and perceptive critique. I was disappointed, for any relationship between the Mincer-Cain comment and my article on "Urban Poverty and Labor Force Participation" [5] is a distant one at best.

In this reply, I shall first respond to the general criticisms of Mincer and Cain and then present some regression results which respond to their specific empirical results. It might be best to begin with what I think was the clearly stated purpose of my paper—namely, "to analyze the significance of patterns of labor force participation for the poverty problem" [5, p. 105]. My study was concerned primarily with "estimating the directions and magnitude of the relation between the overall state of the economy (as approximated by the unemployment rate in urban areas) and the labor force participation rates of the urban poor" [5, p. 105]. An inverse relationship was found between the SMSA unemployment rates and the labor force participation rates of the urban poor, and this relationship was especially significant for non-white married women with husbands present. After posting several caveats, I stated that the basic conclusion which could be drawn from my findings is that "the average non-white urban family with both husband and wife present attempts to lift itself out of poverty . . . by becoming multiple-earner families" [5, p. 117]. Mincer and Cain simply ignore this and assert that "the central and novel conclusion of the study is that the negative response of labor force participation to unemployment is stronger for the poor than for the non-poor, and for non-white married women compared with white married women." Although these results were by-products of the main purpose of my study, they were not the main results of the study and incidentally have not been proven correct or incorrect by the Mincer-Cain comment. I will comment more specifically on this point later in this reply.

Another criticism of my article in the Mincer-Cain comment is that the cross-section regression estimates are not translatable directly into aggregate time series behavior. I agree and explicitly state that "the cross-sectional analyses in this study probably represent the accumulation of labor force adjustments to long-run differences in economic opportunity and social factors in addition to short-run cyclical variations. Thus, it is not entirely legitimate to use the estimates of labor force sensitivity presented in this study to project short-run changes in labor force behavior over the cycle" [5, p. 117]. Possibly, I should have stated this obvious latter point more strongly but the problems of using cross-section regression estimates in analyzing time series behavior have been demonstrated by Kuh [3].

Apparent misunderstandings aside, there is a basic conceptual difference between the Mincer-Cain framework and the approach contained in my article. Mincer and Cain utilize the conventional theory that a worker is presumed to be able to make a rational choice between income and leisure. Utility maximization on the part of the worker is viewed merely as an extension of consumer choice. Mincer and Cain assert that "it is important to keep in mind that labor force participation rates across areas are related to both wage and income variables, and to both normal and transitory components of these variables. The closer the model comes to specifying these variables, the more confidence we have in it."

In the light of their criticisms, it is ironic to

note that the time series regression presented in Table IB utilizes as the independent variable neither a wage nor an income variable but the overall unemployment rate which Mincer and Cain claim is a proxy for the income and substitution variables. It should be noted that the unemployment rate can be used as a proxy for the wage variable (or more precisely, changes in wage rates) over time only if one assumes a stable long-run Phillips Curve relationship. Whether a stable long-run Phillips Curve for the United States exists remains to be demonstrated.

In my paper, I also used the unemployment rate as the primary independent variable, but for different reasons. In my view, it is more useful to regard the unemployment rate as an important variable on its own rather than merely a proxy variable for traditional price theory variables. In the absence of job vacancy data, the unemployment rate becomes a proxy variable for the number of available jobs in an area or the country as a whole. As William Bowen and T. A. Finegan state it, "the unemployment rate can serve as a measure of the probability that an individual job-seeker who is prepared to invest a given amount in search will not be able to find employment within a given period of time" [1]. It is my contention that the key group in my study and in the Mincer-Cain comment, namely poor married non-white women with husbands present, regard jobs per se as the meaningful stimuli for movements into and out of the labor force. I am hypothesizing that a person in poverty compares her present income position which may be zero (being married with spouse present, she is unlikely to be able to collect welfare), against income forthcoming if she can find a job. Even though she is handicapped in her job search by being Negro and most likely unskilled and with little education, a tightening of the labor market, i.e., a lowering of the unemployment rate, enhances her chances of acquiring a job. Information concerning the whereabouts of available jobs is not quite the scarce good it formerly was.

It seems to have been overlooked by Mincer and Cain that I was talking primarily about a very special and unfortunate group in our society—namely the urban non-white poor. Almost 200 years ago, the following words were written: "We hold these truths to be self-evident, that all men are created equal, that they are endowed by their creator with certain unalienable Rights, that among these are Life, Liberty, and the Pursuit of Happiness —That to secure these Rights, Governments are instituted among Men, deriving their just Powers from the Consent of the Governed. . . ." These famous words were written not by a radical New Leftist but by a brilliant slave owner, Thomas Jefferson. Jefferson, however, did not regard Negroes as men in the sense he used the term in the Declaration of Independence. Not even 100 years after a bloody civil war which eliminated formal slavery, can true equality be said to have been accomplished. Mincer and Cain are as aware of this awful phase of American history as I am. But they do not seem to recognize that many poor and unskilled non-whites do not have the opportunity to examine jobs in terms of their pecuniary and nonpecuniary features and then make a decision regarding work effort and leisure. For example, many jobs for which they might be hired by a nondiscriminatory employer pay the going minimum wage. Thus, I contend and will demonstrate later in this reply that the availability of jobs rather than the differences in wages plays a far more important role in explaining the variations in labor force participation rates for *poor* urban non-white married women with husbands present.

On a more specific point, Mincer and Cain assume without even a qualifying footnote that all non-whites are poor and all whites are not poor. This is a "slight" distortion of reality. To be precise, 35 per cent of all non-white persons (8.3 million) fell below the Social Security Administration's poverty line in 1967 and 10 per cent of whites (17.6 million) fell below their poverty line [6, p. 9]. I do not believe that the preceding is a picayune point but on the contrary feel that it makes much of the Mincer-Cain comment irrelevant. Their comment simply is not directed at the poverty problem. In my view, this is the fundamental difficulty with their comment as it pertains to my article.

Mincer and Cain are correct on one point

however. I did not have enough evidence to draw any conclusions about the differences in the negative responses to unemployment of the labor force participation rates of the poor versus the non-poor. They jump on my remarks pertaining to this point as the main conclusion of the article and I simply assert that I did not regard it as such. They go even further by attempting to show that "direct time series observations appear to show that the cyclical labor force response of non-white married women, husband present [a minority of whom are poor incidentally] is weaker than that of white women, and, at times even counter-cyclical." My own empirical results concerning this issue leave me in a state of agnosticism.

The gist of the Mincer-Cain comment is that the regression model used in my study is not completely specified. Of course, they are correct but I admitted this shortcoming in my article [pp. 5, 115–16] and the reasons for it. However, in their time series analysis, Mincer and Cain use the simplest of time series models to explain variations in labor force participation that I have ever seen. For the dependent variable, they use the difference in the labor force participation rates of non-white and white married women with husbands present. However, these participation rates are for only one month in the year (April or March) and are seasonally unadjusted. Since their data refer to only one month per year, there is the distinct possibility that there is a seasonal bias. Furthermore, their independent variable is the unemployment rate for the same month. Their model, not allowing for any lags, simply overlooks the strong possibility that there is actually some lapse of time between changes in labor market conditions and induced changes in labor force participation rates.

I have tried to improve on the form of the Mincer-Cain time series model. I obtained unpublished average quarterly data on the labor force participation rates of non-white and white married women with husbands present. Unfortunately these data are only available from the first quarter of 1959 to the second quarter of 1967. Nonetheless, the results are still of some interest. When the differences in the participation rates are regressed on the average quarterly unemployment rate and a time trend, the coefficient of the unemployment rate was positive (.89) but insignificant (t = 1.3). The R^2 was a whopping .08. When the average quarterly unemployment rate was lagged by one quarter, the coefficient of the unemployment rate became negative and still insignificant. In addition, the R^2 did not show any significant increase. Further experimentations with lagged unemployment rates and distributed lag formulations also failed to improve the explanatory power of the regession model. In addition, if you break the Mincer-Cain time series at year 1959 and run their regression with the unlagged March unemployment rate and a time trend for the years 1959–1967 (the 1967 data are now currently available), the results are very different from their 1947–1966 time series regression. The coefficient of the unemployment rate is still positive (.66) but has a "t" value of .97 and thus cannot be regarded as significantly different from zero. The R^2 falls to .15. Breaking the time series at 1959 is not altogether arbitrary since my study was concerned with cross-section behavior in 1960 and its implications for the future. What do these conflicting time series results mean? As I stated above, I do not know. However, it should be clear that the simple time series results presented by Mincer and Cain cannot be accepted as the final answer to what is obviously a very complicated question.

Regarding the cross-sectional results of Cain's study which are used to demonstrate that the coefficient of the unemployment rate in my study, as it is applied to poor urban non-white married women with husbands present, was "surely biased," a brief coment is required. I admitted in the article [5, p. 116] that the coefficient of the unemployment rate was biased upwards, but not for the reasons stated by Mincer and Cain. To satisfy their criticisms that "an improvement in the model would be to add long-run income and wage variables and other permanent effects so that the measures of short-run or transitory effects can be more nearly isolated," I have

added variables representing the income of non-white males and wage earnings of non-white females for the entire SMSA's under consideration. The initial and "improved" regression results for poor urban non-white married women are presented in Table 1. The N census tracts are those which had a median income less than two-thirds of that for the SMSA as a whole and a non-white population that was over 50 per cent of the total population. The population of the vast majority of these N tracts consisted almost entirely of non-whites. Since this is the group (i.e., non-white married women with husbands present) on which Mincer and Cain concentrate much of their attention, I have limited my analysis to this group. The dependent variable is the labor force participation rate of non-white married women with husbands present in each of the SMSA's.

The addition of the income and wage variables does not alter the coefficient of the unemployment rate in any significant fashion. Once again, it has to be recalled that Cain's work applies to all non-whites in a larger sample of SMSA's, whereas my work pertains only to urban, poor non-whites. For purposes

TABLE 1—RESULTS OF MULTIPLE REGRESSIONS FOR URBAN POOR NON-WHITE MARRIED WOMEN WITH HUSBANDS PRESENT, FOR ALL SMSA's OVER 500,000 IN POPULATION IN 1960

N Tracts	X_1	X_2	X_3	X_4
R^2	b (=net partial regression coefficient) "t" value in parentheses	b	b	b
.48 (n=45)	−3.63 (4.1)	+.87 (2.0)	—	—
.50 (n=42)	−3.74 (3.82)	+1.1 (1.9)	−.31 (.73)	.26 (.75)

Variables
X_1 = the SMSA unemployment rate
X_2 = the industrial structure variable for each SMSA
X_3 = income of nonwhite males in each SMSA
X_4 = income of nonwhite females in each SMSA

of clarification for the reader, this point cannot be overemphasized.

My results would seem to suggest that Negroes should have benefited disproportionately from the most prolonged economic expansion in our history (i.e., the last seven years). There is some evidence that such a phenomenon has occurred. For example, median income of non-white families as a percentage of white family income has increased from 53 per cent in 1961 to 62 in 1967 [6, p. 6]. The percentage of non-white persons below the poverty line fell from 55 in 1961 to 35 in 1967 [6, p. 9]. Whether this improvement in the Negro's economic position has come primarily from an increase in earnings or from an increase in the number of earners in Negro families, it is impossible to ascertain. Undoubtedly, both factors played a role. There is some evidence that the number of earners per Negro family has increased rather rapidly. In 1964, 53.5 per cent of all Negro families had two earners or more. By 1967 this per cent had risen to 58.5. Furthermore, the ratio of median income of Negro families with 3 earners or more to white families with 3 earners or more rose from 48.1 per cent in 1964 to 60.1 per cent in 1967—a very rapid increase in only four years.[1]

I do not want to conclude by implying that economic expansion alone can solve the plight of poor non-whites. The problems created by 300 years of subjugation and discrimination could not possibly be solved by seven plus years of economic expansion. On the contrary, I feel much the same way that the late Dr. Martin Luther King felt when he stated so eloquently,—"Much of the ugly experiences of Negro history have been obscured or forgotten. A society is always eager to cover great misdeeds with a cloak of forgetfulness, but no society can repress an ugly past when the ravages persist into the present. America owes a debt of justice which it has only begun to pay." I would only add that poverty in modern America is so unjust and ugly that all who tolerate it must try to stifle the voice of con-

[1] I would like to thank Dr. Herman Miller of U.S. Census Bureau for providing me these unpublished data.

science within themselves. I am confident there is no disagreement on this latter point between Mincer and Cain and me.

JOSEPH D. MOONEY*

* The author is associate professor of economics at the University of Michigan, Ann Arbor. As a result of a prolonged illness, it was not possible to reply to this comment for several months. Professors Mincer and Cain were gracious enough to allow publication of their comment to be postponed. Thanks are due to Professor Cain for providing some data from his study. Finally, Terrence Kelly made helpful comments on an earlier draft of this reply.

REFERENCES

1. W. G. BOWEN AND T. A. FINEGAN, "Discussion" in R. A. Gordon and M. S. Gordon, ed., 1966, *Prosperity and Unployment*, pp. 113–21.
2. G. CAIN, *Married Women in the Labor Force*. Chicago 1966.
3. E. KUH, "The Validity of Cross-Sectionally Estimated Behavior Equations in Time Series Applications" *Econometrica* No. 27, (1959) pp. 197–214.
4. J. MINCER, "Labor Force Participation of Married Women" in *Aspects of Labor Economics*, National Bureau of Economic Research, Princeton 1962, pp. 63–105.
5. J. MOONEY, "Urban Poverty and Labor Force Participation," *Am. Econ. Rev.*, March 1967, *57*, 104–19.
6. U.S. DEPARTMENT OF COMMERCE, BUREAU OF THE CENSUS, Current Population Reports, Ser. P-23, No. 26, "Recent Trends in Social and Economic Conditions of Negroes in the United States," Washington 1968.

"PRICE" VS. "TARIFF" ELASTICITIES IN INTERNATIONAL TRADE: COMMENT

In his note M. E. Kreinin [1] does not distinguish between a change in the effective rate of protection caused by a movement along a given supply function and a change in the effective rate of protection caused by a shift in the supply function. The result is that all of his main conclusions are wrong.

Kreinin's main argument can be summarized in the following manner. If the nominal tariff levied on the final product j is changed by a greater amount than is the tariff levied on the input i, then the change in the nominal rate on j will be less than the change in the effective rate of protection on j. It is the effective rate of protection and not the nominal rate that determines the production effect of a tariff structure; therefore,

> . . . when elasticities obtained by other means (i.e., without reference to tariff changes) are employed along with nominal tariffs, to estimate the effect of tariff policies (such as multi-lateral reductions or a free trade area in manufactures) on trade flows and welfare, they are likely to underestimate these effects. Only for an across-the-board change which applies to all commodities, will the results be conceptually accurate [1, pp. 893–94].

To demonstrate that the conclusions included in the above quote are wrong, two cases are presented. In one case the effective rate of protection changes by more than the nominal rate, but the price elasticity of import demand will give a correct estimate of the relative change in imports that will occur for any given change in the nominal rate of protection. In the other case, an across-the-board change in tariffs leads to a smaller actual relative change in imports than would be expected on the basis of the price elasticity of import demand.

Assume: (1) Commodities i and j are both imported; (2) Each unit of j produced requires k units of i; and (3) The world prices of i and j are not affected by the tariff policies of the country being examined.

In Figure 1, D is the domestic demand curve for j. S_0 is the domestic supply curve of j when there is no tariff levied on the imports of i. S_1 is the domestic supply curve of j when a tariff of T_i is levied on the imports of i. The vertical distance between S_0 and S_1 is: $\overline{P}_i kT_i$. $\overline{P}_i$ is the world or free trade price of i. $\overline{P}_j$ is the world price of j and $\overline{P}_j(1+T_j)$ is the import price of j when a tariff of T_j is levied on the imports of j. In Figure 1, the value of T_j is chosen so that: $T_j=\overline{P}_i kT_i/\overline{P}_j$. This

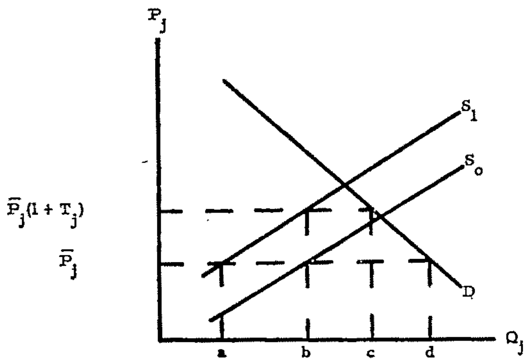


FIGURE 1

means that the domestic production of j is the same when tariffs of T_i and T_j are levied as it is when free trade exists. In Kreinin's terminology, the effective rate of protection is zero. Figure 1 clearly shows that the effective rate of protection is zero whenever the reduced production that occurs at any price due to the leftward shift of the supply curve, which is the result of a tariff being levied on the input, is exactly offset by the increased production that occurs due to the higher domestic price, which is the result of a tariff being levied on the final good.

Suppose that tariffs of T_i and T_j are levied. Further, suppose that the characteristics of the curves S_1 and D are known so that the arc price elasticity of import demand between any pair of prices can be calculated. For instance, between prices $\overline{P}_j(1+T_j)$ and $\overline{P}_j$ the arc elasticity of import demand is:

$$[(c-b)-(d-a)]/[(c-b)+(d-a)]$$
$$\div[\overline{P}_j(1+T_j)-\overline{P}_j]/[\overline{P}_j(1+T_j)+\overline{P}_j].$$

If this elasticity is multiplied by the change in relative import prices that would occur if T_j were reduced to zero, the resulting estimate of the relative increase in imports will equal the relative increase in imports that would occur if the tariff on j was, in fact, reduced to zero and the tariff on i was unaltered. The correct estimate is obtained when only the change in the nominal rate is used, even though the effective rate of protection has changed by more than the nominal rate.[1]

The estimated impact of a reduction of T_j to zero will not change if there is an across-the-board change in tariffs, but the actual relative change in imports will not be the same and it will be less than the estimated relative change. The estimated relative increase in imports will be:

$$[(c-b)-(d-a)]/\tfrac{1}{2}[(c-b)+(d-a)]$$

which is larger than the actual relative change in imports which will be:

$$[(c-b)-(d-b)]/\tfrac{1}{2}[(c-b)+(d-b)].$$

The difference between the estimated relative increase in trade and the actual relative increase in trade occurs because of the incorrect use of the price elasticity of import demand. The elasticity is calculated from knowledge about the curves S_1 and D. The use of the elasticity to estimate the effect of changes in prices is only correct in cases in which these curves remain stable. When the tariff on i is reduced, the supply curve shifts to the right; and the use of the elasticity based upon the position of the initial supply curve will overestimate the change in trade that will actually occur. The conclusion is that the use of price elasticities with nominal tariffs to estimate the effect of tariff policies upon trade flows will be conceptually correct only if changes are made in the nominal rates on the final goods while the tariffs on the inputs remain unchanged. If there is an across-the-board reduction in tariffs, the use of the price elasticity of import demand will be conceptually incorrect and will yield an overestimate of the actual changes that will occur.

It is easier to show that Kreinin's conclusions are wrong than it is to determine what part of his analysis leads to the erroneous conclusions.[2] It appears that the error occurred in the following way. Kreinin observed that T_j has fallen more than T_i,

[1] The change in the nominal rate is T_j. The initial level of the effective rate of protection is: $(T_j-a_{ij}T_i)/(1-a_{ij})$ where $a_{ij}=\overline{P}_ik/\overline{P}_j$. After T_j has been reduced to zero, the effective rate of protection is: $-a_{ij}T_i/(1-a_{ij})$. The change in the effective rate of protection is: $T_j/(1-a_{ij})$. Since $(1-a_{ij})$ is less than 1, T_j is less than $T_j/(1-a_{ij})$.

[2] This paragraph was substantially revised due to the referee's comments.

causing effective protection to fall more than the nominal rate T_j. From this he concluded that tariff elasticities calculated from changes in T_j reflect an underestimate of the "true" price change involved, and thus underestimate the induced fall in domestic production and increase in imports. But the measured price effect of a reduction in nominal tariff T_j alone should cause a movement along the supply curve of import-competing production that is the same as that caused by an equal reduction of $\overline{P}_j$. Any simultaneous fall in T_i shifts the supply curve to the right. Production of import-competing goods then falls less, and imports rise less, than would have been predicted from a price elasticity.

G. DONALD WOOD, Jr.*

* Assistant professor of economics, Georgetown University.

REFERENCE

1. M. E. KREININ, "'Price' vs. 'Tariff' Elasticities in International Trade—A Suggested Reconciliation," *Am. Econ. Rev.*, Sept. 1967, *57*, 891–94.

"PRICE" VS. "TARIFF" ELASTICITIES IN INTERNATIONAL TRADE: REPLY

Professor Wood argues that in relating percentage changes in the quantity imported to percentage changes in the effective tariff rates, I deviated from the traditional concept of elasticity which assumes unchanged input prices. The point is valid and well taken. But, although it nullifies the specific hypothesis advanced in my paper, his argument is directed more at the estimates which I attempted to reconcile, than at the reconciliation itself. Many of the elasticity estimates, however derived, embody changes in input prices, and the use of the elasticity concept is often inappropriate.

It will be recalled that I attempted to reconcile the fact that estimates of import-demand elasticities (for finished manufactures) are higher when obtained from tariff changes than when derived from direct price-quantity relationship. Since import-demand is the difference between domestic demand and supply of the commodity at various prices, any change in input prices would shift the supply function, and distort the elasticity estimate.[1] Incorporating the point made by Wood, we note that neither approach necessarily measures the traditional partial equilibrium elasticity. Under both methods input prices may be expected to change (not necessarily in the same direction as prices of the final product) over the period used in the estimation, either because of tariff changes or for other reasons. Furthermore, even estimates free of this bias are not readily applicable to new situations—e.g., to measure the effect of commercial policies—without prior knowledge of movement of input prices.

Thus, although the a priori formulation suggested in my original note is ruled out, we are still left with a wide range of uncertainty concerning the effect of changes in input prices on the import-demand elasticity and the reconciliation may well lie within this area. How important is this effect is an empirical question, and depends upon circumstances in the period used in the investigation. In an attempt to shed some preliminary light on it, in the case of finished manufactures imported into the United States, I regressed the volume of these imports on: real GNP; the ratio of the import price index to the domestic wholesale price index of finished manufactures; and the domestic wholesale price indexes of crude materials and intermediate materials, where all variables were expressed in logarithm form. In an alternative regression, the import price indexes of the input variables were used. But in both regressions, material input prices failed to show any significant relation to the volume of imports of finished manufactures, or they bore a negative sign. Needless to say, materials are only one input that can shift the domestic supply function; and even they may be more important in other industrial countries than in the United States.

MORDECHAI E. KREININ*

[1] Paradoxically, estimation of domestic demand elasticity requires the assumption of shifts in the supply curve along a constant demand function.

* Professor of economics, Michigan State University.

COUNTRY SIZE AND TRADE PATTERNS: COMMENT

I

In the June 1968 issue of this *Review*, D. B. Keesing has advanced the proposition that small countries are at a disadvantage in the international trade of manufactured goods because their limited national markets restrict the possibilities of exploiting internal and external economies. According to Keesing, this "size effect" would give rise to a positive correlation between per capita exports of individual categories of manufactured goods and population, while there would be a negative correlation between per capita imports and population. In turn, an "income effect" is said to exist if the partial regression coefficient of per capita exports and imports with respect to income per head exceeds unity.

Having tested for the size effect in a cross-section analysis of 31 countries in regard to 40 commodity categories, Keesing claims that "the regression results strongly confirmed the hypothesis. Small countries appeared to experience a comparative disadvantage in most of the important manufacturing industries uncompensated by a comparative advantage in others" [5, pp. 454–55]. He also purports to have found a significant income effect in the relationship between per capita exports (imports) of these manufactured goods and income per head.[1]

It is my contention that the statistical tests suggested by Keesing cannot be used to establish these propositions and, at any rate, his formulation is not appropriate for the problem at hand. I will present evidence on these points and will also formulate and test alternative hypotheses concerning the relationship between country size and trade patterns.

II

Keesing's method provides an example of the "fallacy of composition," i.e., the inappropriateness of arguing from the particular to the general. While the partial elasticity of per capita exports and imports of individual categories of manufactured goods with respect to population is statisticallly significant at the 5 per cent level in 105 of the 160 regressions,[2] this result may be due to influences other than the disadvantages of small countries in manufacturing industries. Thus, the observed positive correlation between per capita exports of particular categories of manufactured goods and population may be explained by the fact that small countries tend to specialize in a few export products and hence only some of these countries—or none of them—will export any one commodity.[3] Specialization, in turn, leads to high per capita imports of various manufactured goods and thus contributes to the observed negative correlation between per capita imports of manufactured goods and population.

To eliminate the influence on the results of the specialization of small countries in a few commodities, I have taken as the dependent variable the per capita exports [equations (1) to (3)] and imports [equations (4) to (6)] of all manufactured goods[4] rather than of individual product categories. In the case of exports, the regression coefficients of the size variable are not statistically significant, irrespective of whether we consider the 31 or the 18 country group. And while there is indication of an income effect in the 31 country group, on closer inspection this is shown to be due to the lack of homogeneity within the sample. Thus, introducing a dummy variable in equation (2) to differentiate between developed and developing countries, the income effect ceases to be statistically significant and there is even less indication of a size effect. Conversely, the size effect, although not the income effect, appears significant in the import regressions [equations (4) to (6)].

[1] Following Chenery [2], the general formula used in the calculations is $log\ (X^{m}/N)_i$ or $log\ (M^{m}/N)_i = log\ a_i + b_{1i} log(Y/N) + b_{2i} log\ N + \varepsilon_i$, where X_i^{m} and M_i^{m} denote the exports and imports of individual categories of manufactured goods, Y refers to gross national product, and N to population.

[2] Regressions have been run separately in regard to the exports and imports of the 40 commodity categories for the entire group of 31 countries and for a subgroup of 18 developed countries.

[3] Keesing raises this possibility but fails to take account of it in deriving his final conclusions.

[4] Defined to include commodity classes 5 to 8 of the Standard International Trade Classification.

REGRESSION EQUATIONS ON THE RELATIONSHIP OF COUNTRY SIZE AND TRADE PATTERNS[1]

Dependent Variable	Number of Countries	Coefficients of Independent Variables Y/N	N	Z	Constant term	R^2	Durbin-Watson
1. X^m/N	31	1.73 (7.3)	0.15 (1.0)		−4.05 (40.6)	0.66	2.22
2.	31	0.79 (1.8)	0.06 (0.4)	0.93 (2.6)	−1.54 (16.9)	0.72	2.31
3.	18	0.36 (0.6)	0.16 (1.2)		0.32 (3.3)	0.11	1.42
4. M^m/N	31	0.94 (10.9)	−0.39 (6.9)		0.84 (22.8)	0.87	2.15
5.	31	0.72 (4.3)	−0.41 (7.2)	0.22 (1.6)	1.44 (40.2)	0.89	2.27
6.	18	0.66 (1.8)	−0.41 (5.2)		1.85 (33.1)	0.67	2.48
7. X^m/X	31	0.60 (3.2)	0.27 (2.2)		−3.33 (41.8)	0.32	1.66
8.	31	0.30 (0.8)	0.24 (2.0)	0.29 (0.9)	−2.54 (31.8)	0.33	1.69
9.	18	−0.46 (0.9)	0.38 (3.6)		−0.41 (5.5)	0.48	1.50
10. M^m/M	31	−0.05 (1.0)	−0.08 (2.1)		0.23 (9.9)	0.15	2.43
11.	31	−0.07 (0.6)	−0.08 (2.1)	0.02 (0.2)	0.28 (11.7)	0.15	2.44
12.	18	0.22 (1.1)	−0.12 (2.8)		−0.44 (14.1)	0.37	3.00
13. X^{sm}/X^m	31	−0.27 (5.7)	−0.09 (2.8)		0.89 (45.3)	0.53	1.61
14.	31	−0.10 (1.2)	−0.07 (2.4)	−0.16 (2.2)	0.44 (24.3)	0.59	1.66
15.	18	−0.13 (0.7)	−0.07 (1.8)		0.40 (13.6)	0.08	1.46

[1] Explanation of symbols: Y = gross national product, N = population, X = total exports, X^m = exports of manufactured goods, X^{sm} = exports of semimanufactures, M = imports, M^m = imports of manufactured goods, Z = dummy variable taking value of 1 for developed countries and zero for developing countries. All variables other than Z are expressed in logs; t values are shown in parentheses.

III

The differences in the results of the export and import regressions can be explained by the existence of a negative correlation between total trade per head and country size[5] which biases downwards the regression coefficients estimated with respect to population in equations (1) to (6). In effect, manufactured exports (imports) per head—the dependent variable in Keesing's formulation—is the product of two ratios: the share of manufactured exports (imports) in total exports (imports) and total exports (imports) per head.[6] These ratios are subject to different influences, and hence the equations cannot provide an appropriate test of the relationship between country size and comparative advantage. For the latter purpose, I have reformulated the regressions by taking the share of the exports (imports) of manufactured goods in total exports (imports) as the dependent variable.

The results shown in equations (7) to (12) give evidence of a size effect in regard to exports as well as for imports in both country groups.[7] There also appears to be an income effect in regard to exports[8] in the 31 country group, but this will again disappear as we introduce a dummy variable in equation (8) to express the degree of development. This modi-

[5] With regard to per capita exports, the following relationship is shown (t-values in parenthesis):

$$\log X/N = 0.28 + \underset{(6.8)}{1.06}\ \log Y/N \underset{(10.9)}{} - \underset{(4.6)}{0.29}\ \log N;\ R^2 = 0.83$$

[6] In symbols, $X^M/N = X^M/X.\ X/N$ and $M^M/N = M^m/M.\ M/N$.

[7] The explanatory power of the regressions is, however, lower than in the case of equations (1) to (6) where the positive correlation between income and exports (imports) raised the value of the correlation coefficients.

[8] Deviations from zero rather than from one are relevant in this case since the dependent variable is now expressed in terms of relative shares while in the previous formulation it was expressed in per capita terms.

fication will, however, reduce only slightly the size effect. On the average, a doubling of the population seems to be accompanied by a one-fourth increase in the share of manufactured goods in total exports while the increase is nearly two-fifths if we consider the developed country group. The magnitude of the size effect, as well as the explanatory power of the regressions, is considerably smaller in the case of imports.

It would appear, then, that small countries are at some disadvantage in the international trade of manufactured goods although the size effect explains only a relatively small part of intercountry variations in trade patterns. The existence of a size effect provides support of Linder's hypothesis that production for domestic markets is a precondition for exporting manufactures [6, p. 87], since large countries are in an advantageous position in producing manufactured goods subject to economies of scale. The results also support the thesis that small countries tend to gain the most from trade liberalization and from regional integration [1].

IV

But manufactured products form a heterogeneous category and, in explaining international specialization, a distinction should be made between semimanufactured goods and finished manufactures. Semimanufactures are mostly standardized commodities produced in single-product firms which require neither a vast array of suppliers of parts and components nor extensive selling and merchandizing efforts in foreign markets. In turn, finished manufactures are differentiated products which may involve vertical as well as horizontal specialization in the production process, and their exportation necessitates considerable effort and expense in selling abroad. Within the manufacturing sector, therefore, we may expect small countries to have a comparative advantage in semimanufactures and a disadvantage in finished manufactures.

This hypothesis, first advanced by Jacques Drèze with regard to Belgium [3, 4], restricts the validity of Linder's proposition to finished manufactures because standardized semiproducts will not generally require the availability of a home market. In fact, we find numerous examples, in countries as diverse as Belgium, Hong Kong, and Portugal, that domestic consumption plays only a supplementary role as market outlet for standardized manufactures. In all such cases, products conforming to certain specifications can be sold in international markets at the going price.

To test this hypothesis, I have introduced the share of semimanufactures in the exports of manufactured goods as a dependent variable, using the same independent variables as beforehand.[9] The results shown in equations (13) to (15) provide evidence of a size effect, indicating that small countries tend to have a comparative advantage in exporting semimanufactures and a disadvantage in exporting finished goods. On the average, the ratio of exports of semimanufactures to all manufactures falls by 7 per cent as population doubles. Again, there is no evidence of an "income effect" if we differentiate between developed and developing countries.

V

I have indicated that information pertaining to individual commodity categories cannot be used to demonstrate the existence of a "size effect" and an "income effect" in international trade, and have also shown that equations incorporating per capita exports (imports) of manufactured goods as a dependent variable do not permit us to derive conclusions concerning the pattern of international specialization. Introducing alternative formulations of the problem, I have offered evidence on the existence of a size effect which creates disadvantages for small countries in the international trade of manufactured goods. Further statistical tests have been provided to indicate that, within the manufacturing sector, small countries tend to have a comparative advantage in semimanufactures and a disadvantage in finished manufactures. These findings support the thesis that small countries are likely to gain the most from trade

[9] In defining semimanufactures and finished manufactures, I relied on the classification scheme suggested by UNCTAD [7].

liberalization and regional integration, because their opportunities will tend to be equalized with those of large nations.

BELA BALASSA*

* The author is professor of political economy at the Johns Hopkins University and consultant to the economics department at the World Bank. This note is based on research supported by the Bank; it should not be taken, however, to represent the Bank's views.

REFERENCES

1. BELA BALASSA, *Trade Liberalization among Industrial Countries: Objectives and Alternatives*, New York 1967, Ch. 7.
2. H. B. CHENERY, "Patterns of Industrial Growth," *Am. Econ. Re.*, Sept. 1960, *50*, 624–54.
3. J. DREZE, "Quelques reflexions sereines sur l'adaptation de l'industrie belge au Marché Commun," *Comptes Rendus des Travaux de la Societé Royale d'Economie Politique de Belgique*, Dec. 1960, 4–26.
4. ———, "Les exportations intra-CEE au 1958 et la position belge," and Jacques Drèze, Jean van Der Rest, and John van Overbeke, "Analyse de la balance commerciale de l'U.E.B.L. en 1958–1959," *Recherches Econ. de Louvain*, 1961 (8), 715–66.
5. D. B. Keesing, "Population and Industrial Development: Some Evidence from Trade Patterns," *Am. Econ. Rev.*, June 1968, *58*, 448–55.
6. S. B. LINDER, *An Essay on Trade and Transformation*. New York, 1961.
7. UNCTAD, *The Definition of Primary Commodities, Semi-manufactures, and Manufactures*, July 2, 1965 (mimeo).

COUNTRY SIZE AND TRADE PATTERNS: REPLY

I would like to welcome and applaud Professor Balassa's note. In all candor, after running my regressions and leaving my research facilities behind, I became so concerned with the "fallacy of composition" problem relating to small-country specialization that, at one stage, I withdrew my article after it had been accepted. I was persuaded to publish it mainly on the grounds that it would stimulate useful research which, at the time, I could not do myself. Hence I am very pleased to see Balassa's note; and all the more so, because it satisfies some of my own further curiosity on the subject. Let me add that the findings presented by G. Hufbauer and by W. Gruber and R. Vernon, at a recent NBER conference [2], and by Chenery and Taylor [1] also help to confirm the existence of important scale and country-size effects in trade.

Balassa's research further underscores the existence of powerful country-size effects while adding to our knowledge of how and where they operate. His results in equations 13 through 15, and the asymmetry of his export and import findings, should help to dispel any suspicion that these effects are due simply to differences in natural resources per capita.

There is another distinct dimension of his and my results. The strength in his export equations of the dummy variable for "development," and the power of this dummy variable to reduce income effects below a significant level, together with my own similar findings, suggest that with "development," countries break through some sort of supply barrier against successful exports of manufactures. This aspect of trade and development certainly begs for further research.

DONALD B. KEESING*

* The author is associate professor of economics at Stanford University.

REFERENCES

1. H. B. CHENERY AND L. TAYLOR, "Development Patterns: Among Countries and Over Time," *Rev. Econ. Stat.*, Nov. 1968, *50*, 391–416.
2. Universities-National Bureau of Economic Research, Conference on Technology and Competition in International Trade, New York, October 11–12, 1968; papers and proceedings to be published.

NATIONAL DEBT IN A NEO-CLASSICAL GROWTH MODEL: COMMENT

I. *Introduction*

In a recent paper in this journal, P. A. Diamond [1] examines the effects of a constant amount of national debt per head on the utility of a citizen. In his growing neoclassical economy with overlapping generations, he shows that individuals living in the asymptotic steady state suffer a loss in utility due to the existence of a national debt if the asymptotic growth rate is less than the return to capital. In this comment we note that (1) there is a tax distribution scheme which can alleviate the burden of the government debt, (2) any additional burden created by an increase in the per capita debt can be nullified to a degree by an appropriate adjustment in the tax distribution, and (3) a tax financing of a social security transfer (a guaranteed per capita income for the retired generation) is superior to debt financing.

II. *Diamond's Model*

Diamond postulates a growing economy with a single good and two factors of production. The production function is represented by $F(K_t, L_t)$ where K_t and L_t respectively represent the capital stock and labor used at time t. The production function is assumed to be linear homogeneous and subject to a diminishing marginal physical product for each factor. Homogeneity of F enables us to write the per capita production function as $y_t=f(k_t)$, where k is the capital per head and $f>0$, $f'>0$ and $f''<0$. The labor supply is assumed to grow exogeneously at rate n and the supply of capital in period t consists of that part of output in the previous period not consumed, i.e., saving. Assuming the factors received their marginal products, we may write the factor price frontier as $w_t=\phi(r_t)$ where w_t is the wage rate and r_t is the return to capital per unit. The frontier has the property that $\phi'(r_t)=-k_t$. Each individual born at the beginning of period t lives for exactly two periods. Let e_{1t} and e_{2t} be the consumption of each of them in the two periods respectively. Diamond assumes that each such individual works only for the first period and is retired for the second. It is assumed that each individual maximizes a utility function $U(e_{1t}, e_{2t})$ subject to the condition that the capitalized value of his consumption stream is equal to his wage, i.e.,

$$e_{1t}+\frac{e_{2t}}{1+r_{t+1}}=w_t.$$

In addition, the function U is assumed to be twice differentiable, lacks satiation, has a diminishing marginal rate of substitution, and its shape is such as to imply that $0<(\partial e_{1t}/\partial w_t)<1$. The saving of each individual during his working life is s_t, which is equal to $e_{2t}(1+r_{t+1})$, and this constitutes his contribution to the supply of capital for the next period in which he is retired. The demand for capital at the beginning of period $t+1$ is given by the marginal physical product schedule of capital for that period. Hence the inverse of the demand function is written as $r_{t+1}=f'(k_{t+1})$, where f' is the marginal product of capital. The entire system of equations describing this growing economy may be written as:

$$(2.1)\qquad e_{1t}=e_1(w_t, r_{t+1}),$$

$$(2.2)\qquad s_t=s(w_t, r_{t+1}),$$

$$(2.3)\qquad e_{2t}=(1+r_{t+1})s(w_t, r_{t+1}),$$

$$(2.4)\qquad e_{1t}+s(w_t, r_{t+1})=w_t,$$

$$(2.5)\qquad w_t=\phi(r_t),\quad \text{where}\quad \phi'(r_t)=-k_t,$$

$$(2.6)\qquad r_{t+1}=f'(k_{t+1})\quad \text{and}$$

$$(2.7)\qquad s_t/(1+n)=k_{t+1}.$$

We thus have seven equations to determine the endogeneous variables r_t, r_{t+1}, w_t, s_t, e_{1t}, e_{2t}, and k_{t+1} in terms of the exogeneous growth rate, n. The first three equations indicate how each individual allocates his wage by maximizing U subject to condition (2.4). Equation (2.5) is the factor price frontier derived from the production function on the assumption that each factor is paid its marginal product. Equation (2.6) and (2.2) represent the demand and supply of capital respectively. Equation (2.7) is an equilibrium condition in the capital market. The variable

s_t is divided by $(1+n)$ in (2.7) because the total saving in period t divided by L_{t+1} is the capital per head available for production in period $t+1$.

Using equations (2.2), (2.5), (2.6), and (2.7), Diamond obtains a first order non-linear difference equation in r, given by,

$$(2.8) \qquad r_{t+1} = f'\left\{\frac{s[\phi(r_t), r_{t+1}]}{1+n}\right\}.$$

Confining attention to the case where the capital market is Walrasian stable for any given w, consideration of the capital market in isolation implies that,

$$(2.9) \quad (dr_{t+1}/dw_t) = \frac{f''(\partial s/\partial w_t)}{1+n-f''(\partial s/\partial r_{t+1})} < 0.$$

The stability of (2.8) requires that,

$$(2.10) \quad 0 < (dr_{t+1}/dr_t) = \frac{-f''(\partial s/\partial w_t)k_t}{1+n-f''(\partial s/\partial r_{t+1})} \leqq 1,$$

the left-hand side of this inequality coming from (2.9) and (2.5). Assuming (2.10) holds, the system reaches a steady state asymptotically at the growth rate, n. Letting the equilibrium interest rate be r, "efficiency" of the steady state path requires that $r \geqq n$. (See [2], [3].) The steady state path with $r=n$ is known as the golden rule path which maximizes the consumption per head among all steady state paths. (See [2], [3].)

III. *Debt and Debt-Induced Taxes*

Diamond introduces an outstanding government debt into the model by assuming that the government may borrow goods during period t by either entering the capital market or by borrowing abroad. The former debt is called internal and the latter debt is called external. Denote the total amount that the government borrows in period t by G_{t+1} and assume that this debt will be repaid with interest at the rate of r_{t+1} during period $t+1$. To describe an asymptotic state in which there exists a positive finite amount of debt per head, assume that $G_{t+1}/L_{t+1}=g$, a constant. The amount the government expends in period t in excess of the amount borrowed is then

$$(3.1) \quad (1+r_t)G_t - (1+n)G_t = (r_t - n)G_t.$$

If this is positive, it must be collected as a tax that is imposed on currently living citizens. If this is negative, the government may dispose of it as a subsidy to its living citizens. If the debt is an internal debt, the debt is nothing more than a device which can be used to redistribute income between the working and retired generations. The tax burden or subsidy required to service the debt can be imposed on the entire population in any desirable manner the government chooses, we suppose. Assume that during period t the government collects the amount $(1-b)(r_t-n)G_t$ from the younger generation and collects the amount $b(r_t-n)G_t$ from the older generation, where $0 \leqq b \leqq 1$. Assume further that the younger generation at time t anticipates that it will be required to pay a tax one period later of the amount $b(r_{t+1}-n)(1+n)G_t$. In anticipation of the future taxes, the younger generation now saves an amount $b(r_{t+1}-n)(1+n)G_t/(1+r_{t+1})$ as a tax provision. After paying current taxes and providing for anticipated taxes, the net wage of each member of the working generation during period t may then be written as,

$$(3.2) \quad w_t^* = w_t - (1-b)(r_t - n)g - s^*,$$

where

$$(3.3) \quad s_t^* = b(r_{t+1} - n)(1+n)g/(1+r_{t+1}).$$

Substituting w_t^* for w_t in equations (2.1) through (2.4), retaining equations (2.5) and (2.6), replacing (2.7) by a new equilibrium condition for the capital market,

$$(2.7') \quad (s_t + s_t^*)/(1+n) = k_{t+1} + hg,$$

where $h=1$ if the debt is internal and $h=0$ if it is external, and adding (3.2) and (3.3), we have nine equations that describe the new system. Diamond analyzes only the case in which $b=0$.

With these new features added, the stability condition for the new system becomes (3.4).

$$(3.4)\quad 0 < (dr_{t+1}/dr_t) = \frac{-f''(\partial s/\partial w_t^*)(k_t + (1-b)g)}{1+n-f''(\partial s/\partial r_{t+1}) - \{1-(\partial s/\partial w_t^*)\}f''g\{(1+n)/(1+r_{t+1})\}^2} < 1.$$

Assuming that the system as now specified is stable, consider the changes in utility of an individual living in the long-run steady state that stem from exogeneous changes in b and g. Letting the long-run equilibrium value of r_t be represented by r, the equation for the return to capital in the steady state is represented by,

$$(3.5)\qquad r = f'\left(\frac{s+s^*}{1+n} - hg\right), \quad \text{where}$$

$$(3.6)\quad s(w^*, r) = s\{[\phi(r) - (1-b)\cdot(r-n)g - s^*], r\}, \quad \text{and}$$

$$(3.7)\qquad s^* = b(r-n)(1+n)g/(1+r).$$

Taking the total derivative of r with respect to b, we find that

$$(3.8)\quad (dr/db) = \frac{(r-n)}{z}\left[\frac{\partial s}{\partial w^*} + \left(1 - \frac{\partial s}{\partial w^*}\right)\cdot\frac{(1+n)}{1+r}\right]f''g, \quad \text{where}$$

$$(3.9)\quad z = (1+n) + (\partial s/\partial w^*)[k + (1-b)g]f'' - (\partial s/\partial r)f'' - b[1-(\partial s/\partial w^*)]gf''\cdot[(1+n)/(1+r)]^2.$$

In view of (3.4), $z>0$. It follows immediately that

$$(3.10)\quad (dr/db) \begin{matrix}<\\=\\>\end{matrix} 0 \quad \text{according as } r \begin{matrix}>\\=\\<\end{matrix} n.$$

Taking the total derivative of U with respect to b and utilizing the first order condition for the individual's equilibrium that $(\partial U/\partial e_1) = (1+r)(\partial U/\partial e_2)$, we obtain

$$(3.11)\quad (dU/db) = (\partial U/\partial e_1)\{(de_1/db) + (1+r)^{-1}(de_2/db)\}.$$

The consumption of each individual in his last period may be written as

$$(3.12)\qquad e_2 = (1+r)(w^* - e_1).$$

Taking the total derivative of e_2 and rearranging the result,

$$(3.13)\quad (de_1/db) + (1+r)^{-1}(de_2/db) = (dw^*/db) + [(w^* - e_1)/(1+r)](dr/db).$$

After calculating the derivative dw^*/db, we may substitute (3.13) into (3.11) to obtain,

$$(3.14)\quad (dU/db) = -(\partial U/\partial e_1)\left[k\frac{(r-n)}{1+r} + (1-b)g + g(b-h)\frac{(1+n)}{1+r}\right]\cdot(dr/db) + \frac{\partial U}{\partial e_1}\frac{(r-n)^2}{1+r}g.$$

It is easy to see from the above that $dU/db = 0$ when $r=n$, and that $dU/db>0$ whenever $r>n$. In inefficient cases where $r<n$, the sign of dU/db is ambiguous. Summarizing,

THEOREM I: Let $r(b, g)$ and $U(b, g)$ be the long-run equilibrium interest rate and the level of utility corresponding to the parameters b and g, where $0 \leqq b \leqq 1$ and $0<g<\infty$. The utility of an individual living in the long run will be increased by imposing a heavier tax burden on the older generation whenever $r(b, g)>n$. In particular, if $r(1, g)>n$, the entire tax burden should be imposed on the older generation if we are to maximize the utility of an individual living in the long run steady state while holding g constant.

Analogously, the change in r with respect to g is,

$$(3.15)\quad \frac{dr}{dg} = -\left[(r-n)\left\{(1-b)\frac{\partial s}{\partial w^*} - \left(1-\frac{\partial s}{\partial w^*}\right)\frac{(1+n)}{1+r}\right\} + h(1+n)\right]\frac{f''}{z}.$$

For the internal debt, where $h=1$, (3.15) may be written as

$$(3.15') \quad \frac{dr}{dg} = -\{(1+n)b+(1-b)(1+r)\} \cdot \left\{(1+r)\frac{\partial s}{\partial w^*} + \left(1-\frac{\partial s}{\partial w^*}\right)(1+n)\right\} \cdot \frac{f''}{(1+r)z},$$

so that unambiguously, $dr/dg>0$, assuming $0\leqq b\leqq 1$. When $h=0$,

$$(3.15'') \quad \frac{dr}{dg} = -b\left(1-\frac{\partial s}{\partial w^*}\right) \cdot \left\{\frac{(1-b)(\partial s/\partial w^*)}{(1-(\partial s/\partial w^*))b} - \frac{(1+n)}{1+r}\right\}\frac{(r-n)f''}{z}.$$

When $r>n$, $\partial s/\partial w^*\geqq b$, then $dr/dg>0$ and when $r<n$, $\partial s/\partial w^*\leqq b$, then $dr/dg<0$. The sign of dr/dg is ambiguous when $(r-n)$ and $(\partial s/\partial w^*)-b$ are of opposite sign.

The change in utility with respect to g, when $h=1$, may be calculated as

$$(3.16) \quad \frac{dU}{dg} = -\frac{\partial U}{\partial e_1}\frac{(r-n)}{(1+r)} \cdot [\{k+(1-b)g\}(dr/dg) + (1+n)b+(1-b)(1+r)],$$

and, when $h=0$, as

$$(3.16') \quad \frac{dU}{dg} = -\frac{\partial U}{\partial e_1}\left[\left\{\frac{(r-n)}{1+r}k + \frac{[(1-b)(1+r)+b(1+r)]}{1+r}g\right\}(dr/dg) + \frac{(r-n)}{1+r}(1-b)(1+r)+(1+n)b\right].$$

In the case of an internal debt, $dU/dg\leqq 0$ according as $r\geqq n$; and when $r<n$, $dU/dg>0$. We may state this result as

THEOREM II: Assuming $0\leqq b\leqq 1$, an individual living in the long-run steady state, where $r>n$, suffers a loss in utility because of the existence of an outstanding internal government debt; but when the economy is on the golden rule path, $r=n$, there is no loss.

Diamond derives Theorem II only for the special case in which $b=0$. Theorem II also holds for an external debt when $r>n$ *and* $\partial s/\partial w^*\geqq b\geqq 0$, but this result may not hold when $b>\partial s/\partial w^*$.

Inasmuch as an increase in g reduces welfare and an increase in b enhances welfare, we may expect that increases in b within the unit interval can be employed to offset the effect of a small increase in the internal debt when $r>n$. If changes in b and g are to be totally offsetting, we require that,

$$(3.17) \quad dU = (\partial U/\partial g)dg + (\partial U/\partial b)db = 0,$$

where $(\partial U/\partial g)$ and $(\partial U/\partial b)$ are given by equations (3.14), (3.16) and (3.16′) and represent the partial derivatives holding b and g respectively constant. Solving for db/dg, in the case of the internal debt,

$$(3.18) \quad (db/dg) = \frac{(1+n)b+(1+r)(1-b)}{(r-n)g}, \quad \text{where } r>n.$$

It can also be verified that along the locus of b and g for which (3.17) and (3.18) hold, the values of w^* and r (and hence e_1 and e_2) remain constant. Summarizing, we have the following result. Let the initial long-run steady state position be denoted by r', $w^{*\prime}$ and U' with its parameter values given by b' and g'. Then the changes in b and g in accordance with (3.18) will not only leave the utility unchanged but also the r' and $w^{*\prime}$. From (3.18) the locus of b and g in the parameter space with this property can be represented by

$$(3.19) \quad g = \frac{b'g'}{(1+r)-(r'-n)b},$$

where $r'>n$ and $0\leqq b\leqq 1$. Hence we have the following theorem.

THEOREM III: Let b' and g' be the parameter values corresponding to the initial equilibrium. Let $0 \leq b' < 1$. Then the effect of a small increase in the per capita debt may be offset by an appropriate increase in b so as to nullify the additional burden created by an additional per capita debt. Such an offsetting parameter adjustment is represented by (3.18) and (3.19).

IV. *Tax versus Loan Finance; a Guaranteed Social Security Payment*

The trend in the advanced western countries is toward a minimum guaranteed retirement income to each member of the retired generation. Within the context of the model we may analyze the long-run differential impact of loan versus tax finance of such a social security transfer.

Suppose that the government raises the required fund either by borrowing in the capital market and/or by taxing the younger working generation during period t and transfers the proceeds to the older generation of period t. We shall show that tax financing is unambiguously superior to debt financing in performing such a transfer.

Let T_t be the amount of tax imposed on the younger generation in period t. Let G_t be the amount of internally borrowed goods in period t. Letting τ be the guaranteed minimum retired income per member of the older generation, the total transfer required in period t is τL_{t-1}. Hence we must have,

$$T_t + G_t = \tau L_{t-1}. \tag{4.1}$$

Then the tax per head of the younger generation can be written as

$$\pi_t = \frac{\tau}{1+n} - g_t, \tag{4.2}$$

where $g_t = (G_t/L_t)$. From (4.2) we see that π_t and g_t are clearly substitutable parameters for any given τ. The total transfer that the older generation receives may be divided into two parts:

$$\tau L_{t-1} = V_t + (1 + r_t)G_{t-1}, \tag{4.3}$$

where V_t is the amount in excess of the debt payment, $(1+r_t)G_{t-1}$, so as to make the total transfer received per member of the older generation equal to the guaranteed τ. Hence each oldster's additional transfer receipt in excess of the receipt from debt payment is,

$$(V_t/L_{t-1}) = \tau - g_{t-1}(1 + r_t). \tag{4.4}$$

In view of the specification given above, the net wage constraint for a member of the younger generation is now,

$$\begin{aligned} w_t^* &= w_t - \pi_t + \frac{V_{t+1}}{L_t(1 + r_{t+1})} \\ &= w_t - \frac{\tau(r_{t+1} - n)}{(1 + r_{t+1})(1 + n)} . \end{aligned} \tag{4.5}$$

The capital market equilibrium in period t is now expressed as,

$$r_{t+1} = f' \left\{ \frac{s(w_t^*, r_{t+1})}{1+n} - g_t \right\} . \tag{4.6}$$

The stability condition for the system specified becomes, (4.7).

$$0 < (dr_{t+1}/dr_t) = \frac{-f''(\partial s/\partial w_t^*)k_t}{1 + n - f''(\partial s/\partial r_{t+1}) + f''(\partial s/\partial w_t^*)\tau(1 + r_{t+1})^{-2}} < 1. \tag{4.7}$$

$$\frac{dr}{dg} = \frac{-f''(1 + n)}{1 + n + f''\{k + \tau(1 + r)^{-2}\}(\partial s/\partial w^*) - f''(\partial s/\partial r)} \tag{4.8}$$

Assuming (4.7) holds, from (4.6) the asymptotic interest rate is now represented by,

$$(4.6') \qquad r = f'\left\{\frac{s(w^*, r)}{1+n} - g\right\}.$$

Taking the total derivative of r with respect to g in (4.6′), we obtain (4.8) which is positive from (4.7)

Noticing that π and g are substitutable parameters for any given τ, as in (4.2), the substitution of debt financing for tax-financing increases the long-run interest rate. The corresponding change in the long-run utility becomes,

$$\begin{aligned}(4.9) \quad (dU/dg) &= (\partial U/\partial e_1)(dr/dg)(1+r)^{-1} \\ &\cdot\left\{k(n-r) + g(1+n) - (1+r)^{-1}\right\}.\end{aligned}$$

Assuming that the guaranteed minimum retirement income is larger than the amount the older generation receives from the debt holding (the amount of transfer in the absence of social security system), we have from (4.4),

$$(4.10) \qquad \frac{\tau}{1+r} > g(1+n).$$

Therefore in the efficient steady state where $r \geqq n$, dU/dg is always negative. Hence debt financing of social security transfers is inferior to tax financing in efficient cases.

G. O. Bierwag, M. A. Grove, and Chulsoon Khang*

* The authors are members of the Department of Economics, University of Oregon.

REFERENCES

1. P. A. Diamond, "National Debt in a Neoclassical Growth Model," *Am. Econ. Rev.*, Dec. 1965, *55*, 1126–50.
2. E. S. Phelps, "Second Essay on the Golden Rule of Accumulation," *Am. Econ. Rev.*, Sept. 1965, *55*, 793–814.
3. R. M. Solow, "Comment," *Rev. Econ. Stud.*, June 1962 ,*29*, 255–57.

MIXED PUBLIC AND PRIVATE FINANCING OF EDUCATION: COMMENT

In a recent article, Mark V. Pauly deals with mixed public and private financing of education [4]. The need for mixed financing arises because education creates both private and public benefits. It is Pauly's aim to discover a solution to the financing problem which allows a community to achieve the socially desirable level of educational output.

In this note, it will be argued that Pauly's conceptual treatment of education should be generalized. Pauly does not separate public from private output; he sees education as a good which yields utility to the person receiving instruction while entering at the same time into the utility functions of all other members in the community. The alternate formulation suggested in this note treats educational output as consisting of two components, a public and a private one, which are produced under conditions of joint supply and which enter separately into people's utility functions. As a result, we find that an efficient solution to the financing problem requires the achievement not only of the optimal level but also of the optimal mix of educational output.

I. *Education as a Case of Joint Supply*

Many authors have drawn attention to the dual nature of educational output. While the educational process yields important benefits which are enjoyed exclusively by the child attending school and by his family, it also creates benefits which accrue to other members in the community. A brief discussion will clarify the character of such external benefits. One type—that associated with general literacy—is often mentioned. Widespread literacy greatly facilitates the transaction of community business and the communication of matters affecting the public interest. A second type of external benefit, even more fundamental to community life, is of an ideological nature. Education helps to transmit a common set of values and attitudes thus contributing to the creation of those preconditions which are necessary to

make a stable economic and social life possible.

Public and private benefits are the result of a joint production process. A child who learns to read and write is acquiring skills which have a private as well as a public use. It is important to realize that both types of public benefits are involved. The child is not only becoming a literate citizen, he is also acquiring a set of attitudes toward matters which affect the community. Books used in the teaching of reading and writing are not free of values and they are bound to affect the child's outlook and his reactions to public issues in later life. This will be true to an even stronger degree for materials used in the teaching of other kinds of knowledge such as history, geography, and languages.

Once we see public benefits in this broad perspective, it becomes clear that a given amount of economic resources can produce varying combinations of public and private output. In particular, the "ideological" content of courses and books can be adjusted in a way which affects the resulting proportion of public and private benefits. On the other hand, the discussion also implies that separate production is not possible. We are dealing with a case of joint supply, and Marshall's well-known example—the production of meat and hides—illustrates the problem [3, pp. 388–91]. Like the breeder who cannot produce meat without hides and hides without meat, the educator cannot create purely private or purely public benefits.[1] While books and courses stressing private skills contain elements affecting public output, courses which specialize in "civics," i.e., citizenship education per se, lead to the acquisition of verbal and deductive skills which enhance private earning power.

II. *Implications for Efficient Allocation and Financing*

We now extend Pauly's model to allow for joint supply. Let us assume that we are dealing with a community of m families who have $n+1$ private goods (including private educational output) available to them. The utility function of the ith family may be written as

$$(1) \quad U^i = U^i(X_{1i}, \cdots, X_{ni}, X_{oi}; \; X_{p1}, \cdots, X_{pi}, \cdots, X_{pm}).$$

It differs from the one used by Pauly in two respects. X_{oi}, the private education of family i, enters as an additional private good into the function, while the symbol X_p now refers only to public benefits. It should be noted that family i consumes X_{pi}, the public output associated with its own education, in the same manner in which it consumes public output associated with the education of others.

Any family's education involves the joint supply of a public good, X_p, and a private good, X_o. For Pareto optimality, the following conditions must hold for the public and private output of family j:

$$(2) \qquad \sum_{i=1}^{m} u_{pj}^{i}/u_{r}^{i} = F_{pj}/F_r$$

$$(3) \qquad u_o^j/u_r^j = F_{oj}/F_r$$

where the terms on the lefthand side of the equalities represent ratios of marginal utilities while those on the right-hand side stand for ratios of marginal costs. The subscript r indicates a (private) numeraire good.

It is important to note that our formulation implies the existence of a trade-off at the margin between private and public output. Dividing (2) by (3), we get an equality which makes the trade-off explicit:

$$(4) \qquad \left(\sum_{i=1}^{m} u_{pj}^{i}/u_{r}^{i}\right) u_r^j/u_o^j = F_{pj}/F_{oj}.$$

What are the implications for the financing problem? This question can best be dealt with by examining the operation of a general subsidy program. Let us assume that we have a situation in which preferences are revealed or that we have some mechanism for determining a public subsidy which covers public benefits. Each family receives the "correct" subsidy payment from the community. The subsidy must be spent on edu-

[1] The analogy with Marshall's example, while instructive, is not complete. Meat and hides can be resold, but services such as education cannot be retraded. For a general treatment of externality and joint supply, see [1].

cation but each family remains free to choose its own private school. The private educational sector is competitive.[2]

Under the circumstances outlined, the family which maximizes utility will allocate its expenditure on education, including the subsidy, on the basis of its private rate of substitution. Rewriting (4) we have:

$$(5)\qquad \left(\sum_{i\neq j}^{m} u^{i}_{pj}/u^{i}_{r} + u^{j}_{pj}/u^{j}_{r}\right)u^{j}_{r}/u^{j}_{s} = F_{pj}/F_{oj}.$$

Since the valuation which other members in the community put on X_{pj} does not enter into the family's utility function, only the right-hand term in the bracket is relevant to its decisions. The family will therefore substitute private for public education and buy less public education than is socially desirable. It makes this adjustment by choosing a private school offering the product mix which it desires.

It may be objected that a less general subsidy, one carefully designed to support only certain types of education, can be used as a policy instrument to achieve the proper mix of private and public benefits. While such an argument is theoretically correct, the practical difficulties involved should not be overlooked. We have pointed out that the content of books, other teaching materials, and courses as well as the course mix can be adjusted to yield different proportions of public and private output. This suggests an optimal grant structure of considerable complexity.[3] Because of the problems and costs which must attend implementation of any such scheme, a public school system may offer a better solution for attaining the optimal mix than a privately operated one which is partially supported by the community.

In conclusion we may note that the problem of product mix disappears if public and private educational output is produced in fixed proportions. In that case, any subsidy to private education will raise public education without substitution between the two. While the assumption of fixed proportions simplifies the analysis, it is not in accordance with the observed nature of the educational process. Future work on the financing of education should take the trade-off between public and private benefits into account.

WALTER HETTICH*

* The author is assistant professor of Economics at Queen's University in Kingston and research associate at the Canada Council in Ottawa, Canada. He would like to thank his colleagues, Neil Swan and Dan Usher, of Queen's University, and the referee for their comments on an earlier draft.

REFERENCES

1. J. M. BUCHANAN, "Joint Supply, Externality and Optimality," *Economica*, Nov. 1966, *29*, 404–15.
2. MILTON FRIEDMAN, *Capitalism and Freedom*. Chicago 1963.
3. ALFRED MARSHALL, *Principles of Economics*, Ninth (Variorum) Ed. New York 1961.
4. M. V. PAULY, "Mixed Public and Private Financing of Education: Efficiency and Feasibility," *Am. Econ. Rev.*, March 1967, *57*, 120–30.

MIXED PUBLIC AND PRIVATE FINANCING OF EDUCATION: REPLY

The major point of Mr. Hettich's comment is fundamentally correct. If my model is extended to permit variability in the educational mix, then this variability needs to be considered in defining optimal provision of education. Global optimality requires provision of the optimal quantity of the public good, ***optimally mixed***. For public policy purposes, however, the import of Hettich's analysis may be

[2] The best-known proposal for public subsidies along this line is the one put forward by Milton Friedman [2, pp. 85–107].

[3] The primary difficulty arises from what we have called the "ideological" nature of public benefits. Since adjustment of the ideological content is necessary, we need rather far-reaching control over the curriculum. Economists have at times wondered why public schools are an almost universal phenomenon in today's world and why private but subsidized alternatives are rarely contemplated even in countries which rely heavily on private enterprise. One important answer would seem to lie in the difficulty of adjusting ideological content in a system of private schools

less than he implies. This is so for at least two reasons:

1. It may be that the "community-oriented" content of courses of study which would be chosen by parents would not be smaller than the optimal community-oriented content of such courses. The amount of "common values needed to make stable economic and social life possible" which is contained in private-school courses in history or geography does not appear to be appreciably less than that in similar courses in public schools. General literacy-oriented courses, on the other hand, do not appear to be susceptible to much variation in mix; there are not distinct public and private language. In short, even when variability is possible, Hettich's equation (5) will be unsatisfied by private educational decisions only if there are *marginal* external benefits arising from an increase in the community-oriented component of the educational mix. This may not be the case.

2. Even if there are marginal benefits from altering the privately chosen mix, this may only mean, as Hettich notes, that a subsidy for specific types of schooling would be needed for optimality. Hettich cities "practical difficulties" as an objection to this. But such administrative costs are problems of second-best optimizing, and they are something which no one knows anything about. It seems foolhardy to attempt any generalizations.

An alternative to specific subsidy schemes, and an alternative which is currently employed in most countries which have subsidized or unsubsidized private schools, is regulation of the educational mix. Optimality would only require that the courses or the content of courses, or the qualifications or education of teachers, be specified. Indeed, in many states, the curricula and syllabi of many courses are specified for both public and private schools.

In any case optimality in the *provision* of education (or of any public good) raises problems which are conceptually separate from those concerned with the manner of *production* of the good. It is possible to have public provision without public production, public production without public provision, or a combination of both. There may be reasons for the socialization of the *production* of some kinds of education for some persons within a public school system, but the possible existence of externalities arising from variability in the educational mix is not one of them.

MARK V. PAULY*

* The author is assistant professor of economics, Northwestern University.

NOTES

New Quarterly Journal

Public Policy, previously published as an annual by the John Fitzgerald Kennedy School of Government at Harvard University, will be issued quarterly as of Fall, 1969. The shift in time will permit the publication of professional discussions on public policy issues as they occur. The journal is soliciting a wide range of articles and will attempt to review articles quickly.

Editorial communications should be sent to The Editors, PUBLIC POLICY, Littauer Center 222, Harvard University, Cambridge, Massachusetts 02138.

Announcements

The Journal of Economic Theory will publish original articles on economic theory and related mathematical techniques. The Journal will be published quarterly commencing in the early part of 1969. The editor is Professor Karl Shell, Department of Economics, University of Pennsylvania, Philadelphia, Pennsylvania 19104. The Journal is published by Academic Press, Incorporated, 111 Fifth Avenue, New York, New York 10003.

The Economics Institute is sponsored by the American Economic Association for foreign students of economics and agricultural economics beginning graduate work in the United States.

12th session to be held from June 18–August 20, 1969 at the University of Colorado. Information and application forms may be obtained from the Director of the Economics Institute, University of Colorado, Boulder, Colorado 80302.

The National Gaming Council, 1717 Massachusetts Ave., N.W., will be publishing a quarterly newsletter devoted to news about gaming (model building) dealing with techniques, new models, etc. The newsletter would like to hear from Economists who are working in this field and who would like to use the newsletter as a means of communicating with others in the field about their work. Announcements about meetings, and plays will also be accepted. Please send information to: Peter House, Editor, The National Gaming Council Newsletter, Urban Systems Simulations, 1717 Massachusetts Ave., N.W., Washington, D.C. 20036.

The International Society for the History of the Behavioral and Social Sciences (ISHOBSS) will hold a meeting May 9, 10, and 11, 1969 at Princeton University. For information about submission of papers and/or attendance, please contact Professor Julian Jaynes, Department of Psychology, Princeton University, Princeton, New Jersey 08540.

The Survey Research Center of The University of Michigan will hold a Summer Institute in Survey Research Techniques for the twenty-second consecutive year.

The Institute is designed to meet some of the educational and training needs of men and women engaged in business and governmental research and other statistical work, and graduate students and university instructors interested in quantitative research in the social sciences.

The 1969 Institute will be presented in two four-week sessions, the first from June 30 to July 25 and the second from July 28 to August 22. These two sessions may be taken independently or successively.

For further information please write to the Executive Secretary to the Director, Survey Research Center, The University of Michigan, P.O. Box 1248, Ann Arbor, Michigan 48106.

Deaths

Oscar L. Altman, treasurer of the International Monetary Fund, December 23, 1968.

Aaron Hamilton Chute, professor emeritus of the department of marketing administration, The University of Texas, August 19, 1968.

Gerhard Colm, Chevy Chase, Maryland, December 25, 1968.

George W. Eckelberry, professor emeritus, accounting, The Ohio State University, October 30, 1968.

Clarence E. McNeill, professor emeritus, University of Nebraska, January 3, 1969.

Orville K. Thompson, July 23, 1968.

Retirements

Burton R. Morley, professor of management, School of Commerce and Business Administration, University of Alabama, August 1968.

John Arch White, professor of accounting, The University of Texas, May 31, 1968.

Visiting Foreign Scholars

Dimitrije Dimitrijevic, director, National Bank of Yugoslavia: visiting professor of economics, Florida State University, fall 1969.

Etienne Kirschen, University of Brussels: visiting professor of economics, Drew University program on the European economic community, Drew University.

Suk Hyun Lim: College of Business lecturer in management, School of Commerce and Business Administration, University of Alabama, 1968–69.

Jean Noiret, Nederlands Opleidings Instituut voor het Buitenland (NOIB), Breukelen, The Netherlands: European exchange professor, University of Oregon, College of Business Administration, 1968–69.

Edward Sieper, Massey University, New Zealand: visiting associate professor of economics, University of Rochester, 1968–69.

Keith Sloane, senior lecturer, School of General Studies, The Australian National University, Canberra: lecturer in economics, department of economics, The University of Texas, spring 1969.

Francisco A. Trevino, Monterey Institute of Technology, Mexico: visiting assistant professor of economics, University of Arizona, 1968–69.

John Wadsworth, chief economist, Midland Bank, London, England: visiting professor of banking, University of Georgia, spring 1969.

Promotions

Charles I. Bartfeld: associate professor, management sciences; and director ADPS program, School of Business Administration, The American University, 1968–69.

Carl E. Basham: professor of business administration, West Liberty State College.

Richard A. Bilas: professor of economics, Georgia State College, 1968–69.

Walter P. Blass: director, Revenue Requirements Studies, New York Telephone Company, December, 1968.

Harold L. Buma, Bank of America: manager, economics department, Wells Fargo Bank, San Francisco, California.

Fawzi G. Dimian: assistant professor, department of economics and commerce, Simon Fraser University.

Fred R. Glahe: associate professor of economics, University of Colorado.

Robert F. Godfrey: associate professor of accounting, West Liberty State College.

Loyal M. Hartman: professor of economics, Colorado State University.

Craig G. Johnson: assistant professor, Columbia University Graduate School of Business, Fall 1968.

Joseph H. Levine: assistant professor, Business Law, School of Business Administration, The American University, 1968–69.

James H. Lewis: professor of economics, Colorado State University.

Albert G. Madsen: associate professor of economics, Colorado State University.

John C. McManus: special lecturer in economics, University of Toronto.

J. Robert Moore: assistant professor of economics, Millersville State College.

Raghu Nath: associate professor of business administration, University of Pittsburgh.

David M. Nowlan: associate professor, University of Toronto.

Martin Pfaff: associate professor and director, Master of Business Administration Program, School of Business Administration, The American University, 1968–69.

Frederick B. Putney: assistant professor, Columbia University Graduate School of Business, fall 1968.

Allen B. Rosenberg: associate professor of economics, West Liberty State College.

Melvin D. Skold: associate professor of economics, Colorado State University.

Michael H. Spiro: associate professor of business administration and economics, University of Pittsburgh.

Beryl W. Sprinkel: senior vice president and economist, Harris Trust and Savings Bank.

Bernard Udis: professor of economics, University of Colorado.

Milton D. Vaughn: associate professor of economics, Colorado State University.

Porter K. Wheeler: assistant professor of economics, Wesleyan University, January 1968.

Thomas A. Wilson: professor of economics, University of Toronto.

Ronald A. Wykstra: associate professor of economics, Colorado State University.

Administrative Appointments

Forrest C. Blodgett: chairman, department of economics, Linfield College.

Floyd S. Brandt: chairman, department of management, The University of Texas.

Robert K. Brown: associate dean, graduate school of business, University of Pittsburgh.

Harold K. Charlesworth:: associate dean for extension, College of Business and Economics, and director, Office of Development Services and Business Research, University of Kentucky.

Donald Cole: director, Semester on the European Economic Community, Drew University.

Jon Cunnyngham: chairman, department of economics, The Ohio State University.

Max E. Fletcher: chairman, department of economics, College of Business Administration, University of Idaho.

Raymond R. Gamby: professor and head of department of business administration, Institute of Administration, Ahmadu Bello University, Zaria, Nigeria, October 1968.

Robert E. Georges: assistant dean and director of Undergraduate Business Administration Programs, The Ohio State University.

Karl D. Gregory, Wayne State University: assistant to the Chancellor for Urban Affairs and associate professor of economics, Office of the Chancellor, Oakland University, 1968–69.

M. Khalil Hamid: chairman, department of economics, Millersville State College.

Donald R. Herzog: professor of business and coordinator of graduate programs in business, Chico State College, September 1968.

Franklyn D. Holzman: chairman, department of economics, Tufts University.

Gaylord A. Jentz: chairman, department of general business, The University of Texas, September 1968.

James R. Kay: chairman, department of finance, The University of Texas, July 1968.

Jack Nall: director, management program for executives, Graduate School of Business, University of Pittsburgh.

Walter J. Primeaux, Jr.: professor of business economics and head of department of business administration, McNeese State College, September 1968.

Allen M. Sievers: chairman, department of economics, College of Letters and Science, University of Utah.

Edward K. Smith, deputy assistant secretary of commerce: chairman, department of economics, Colorado State University.

W. Tate Whitman: chairman, department of economics, Emory University.

James G. Yoho, Duke University: head, department of forestry, Mississippi State University.

H. J. Zoffer, dean, Graduate School of Business, University of Pittsburgh.

New Appointments

Robert C. Anderson: assistant professor, Tufts University, 1968–69.

Willis E. Anthony, Marketing Economics Division, USDA: assistant professor and extension economist marketing, Agricultural Economics and Extension Service, University of Minnesota, fall 1968.

Robert D. Auerbach: instructor, University of Illinois.

Richard C. Barth: assistant professor of economics, Colorado State University.

Gary Becker, Columbia University: Ford Foundation visiting professor, University of Chicago.

Huntley H. Biggs: temporary assistant professor of economics, Colorado State University.

Richard M. Bird, Harvard University: associate professor, department of political economy, University of Toronto.

Donald T. Bisesti: economist, New York State Department of Commerce, division of Economic Research and Statistics, Albany, New York.

William A. Brodk, University of California Operations Research Center: assistant professor of economics, University of Rochester, September 1968.

Kirk L. Burns, management consultant: associate professor, School of Business Administration, The American University, 1968–69.

Miltiades Chacholiades: associate professor of economics, Georgia State College, 1968–69.

Juei Ming Cheng: assistant professor of economics, Georgia State College, 1968–69.

Richard B. Clemmer: assistant professor of economics, Georgia State College, 1968–69.

Gail C. A. Cook, Wayne State University: assistant professor, University of Toronto.

Gilbert L. Crouse, Purdue University: lecturer in economics, State University of New York.

Robert L. Darcy: professor of economics, Colorado State University.

Frank de Leeuw: staff member, the Urban Institute.

Forest J. Denman: instructor in economics, Georgia State College, 1968–69.

William C. Dulin: associate professor and director, Business Management, School of Business Administration, The American University, 1968–69.

Liang-Shing Fan: associate professor of economics, Colorado State University.

Rashi Fein, Brookings Institution: professor of economics of medicine, Harvard Medical School and John Fitzgerald Kennedy School of Government, Harvard University, 1968–69.

Victor R. Fuchs: professor of economics, Graduate Center, Mount Sinai School of Medicine, City University of New York, 1968–69.

Frank W. Gill: visiting associate professor of accounting, University of Oregon, 1968–69.

Melvin N. Greenball, University of California: associate professor, accounting, The Ohio State University, 1968–69.

Peter B. Griffin: instructor in economics, Whitman College.

Irvin Grossack: associate professor, business economics and public policy, Graduate School of Business, Indiana University.

Joan Haworth, University of Oregon: joint appointment as assistant professor of economics and consultant to the Computing Center, Florida State University, winter 1969.

William Kempey, New York University: instructor, Hofstra University.

Howard P. Kitt, Columbia University: instructor, Hofstra University.

Gangadhar Kori, Southern Illinois University: associate professor, Business Management, School of Business Administration, The American University, 1968–69.

Mordechai E. Lando: economist, Center for Naval Analyses of the University of Rochester.

Charles Richard Long: assistant professor of economics, Georgia State College, 1968–69.

Sven Lundstedt, Case Western Reserve University: associate professor, research, The Ohio State University, 1968–69.

Robert A. Meier: visiting professor, department of accounting, School of Business Administration, San Diego State College, 1968–69.

James C. Miller III: assistant professor of economics, Georgia State College, 1968–69.

John H. Munro, University of British Columbia, Vancouver: associate professor, University of Toronto.

Alan P. Murray, chief economist, Joint Committee on Internal Revenue Taxation: senior economist, First National City Bank.

Seiji Naya, National Bureau of Economic Research: associate professor of economics, University of Hawaii, spring 1969.

Daniel H. Newlon: assistant professor of economics, State University of New York.

Verne E. Odmark, San Diego State Uni-

versity: visiting professor, accounting, The Ohio State University.

Edgar O. Olsen: economics department, Rand Corporation, January 1969.

Mancur L. Olson, Jr.: associate professor, department of economics, University of Maryland. February 1969.

Terutomo Ozawa: assistant professor of economics, Colorado State University.

Robert Perloff: professor of business administration, University of Pittsburgh, July 1969.

E. Grosvenor Plowman, University of Maine: visiting professor in transportation and business logistics, The Ohio State University, January-April, 1969.

Rolando E. Polli: assistant professor of business administration, University of Pittsburgh.

Ronald L. Racster, University of Southern California: assistant professor, finance, The Ohio State University, 1968–69.

Kenneth J. Rea, University of Saskatchewan: associate professor, University of Toronto.

Larry P. Ritzman, Michigan State University: assistant professor, management sciences, The Ohio State University, 1968–69.

Kenn Rogers, City College: professor, Business Administration, The American University, 1968–69.

Robert W. Rosen: visiting professor, department of management, School of Business Administration, San Diego State College, September 1968.

Mahmoud Sakbani: instructor, Tate University of New York.

Theodore P. Scheinman: instructor, West Liberty State College, 1968–69.

Robert E. Schlenk: supervisor of Socioeconomic Research, State of Michigan, department of Social Services.

Michael E. Scorgie: visiting lecturer, accounting, The Ohio State University.

Lawrence Senesh: professor of economics, University of Colorado, 1968–69.

Jay Siegel, Stanford University: assistant professor, University of Toronto.

Stephen D. Slingsby: assistant professor public administration, The Ohio State University.

Edward K. Smith: professor of economics, Colorado State University.

Manuel Smith: assistant professor of economic development for the Division of Man-Environment Relations, The Pennsylvania State University.

Vernon L. Smith: professor of economics, University of Massachusetts.

David J. Smyth: professor of economics, State University of New York at Buffalo.

Donald M. Sorensen: extension assistant professor of economics, Colorado State University.

Zane A. Spindler, Simon-Fraser University: economics department, Michigan State University.

G. Fred Starner: instructor of economics, Drew University.

Houston Stokes: assistant professor, University of Illinois.

V. B. Subramanian: instructor in economics and accounting, Eastern College.

Philip B. Teets: assistant professor, University of Illinois.

Simón Teitel, United Nations, Industrial Development Organization: advisor, office of the program advisor, Inter-American Development Bank.

Herbert M. Thompson, Jr.: assistant professor of economic research, University of Alaska, 1968–69.

Eric J. Toder: assistant professor, Tufts University, 1968–69.

David L. Vinje: instructor in economics, Whitman College, 1968–69.

Carole A. Walker: instructor in business administration, University of Pittsburgh.

Richard G. Walsh: professor of economics, Colorado State University.

Barbara Ward: professor of international economic development, Columbia University Graduate School of Business, fall 1968.

Leonard Waverman, Massachusetts Institute of Technology: assistant professor, University of Toronto.

Robert E. Whedbee: faculty affiliate, department of economics, Colorado State University.

Joseph Yaney: assistant professor, management sciences, The Ohio State University.

Dieter K. Zschock: acting director, eco-

nomic research bureau, State University of New York.

Leaves for Special Appointments

Leslie P. Anderson, University of Oregon: visiting professor, Harvard University, 1968–69.

Michael E. Borus, Michigan State University: economic Policy Fellow, The Brookings Institution and U.S. Department of Labor.

Norris C. Clement, San Diego State College:: Fulbright visiting assistant professor, Universidad Agraria Del Norte-Lambayeque, Peru; Unversidad Nacional de Trujillo, Peru.

Paul Fisher, chief, International Staff, Social Security Administration: senior research economist, International Labor office, Geneva, Switzerland.

Marshall Geer III, University of Colorado: visiting assistant professor, Universidad Autonoma de Guadalajara, Mexico, 1968–69.

James P. Houck, Jr., University of Minnesota: visiting associate professor of economics, Agricultural and Technical State University of North Carolina, spring 1968.

H. Michael Mann, Boston College: special economic advisor to the Assistant Attorney General in charge of antitrust, Department of Justice, Washington, D.C.

Minnie C. Miles, School of Commerce and Business Administration, University of Alabama: visiting professor of management, Pusan National University, spring 1969.

Thomas G. Moore, Michigan State University: senior staff economist, Council of Economic Advisers, Washington, D.C.

Bertram Silverman, Hofstra University: Ford Foundation Research Award for Cuban Study.

James Witte, Indiana University: visiting professor of economics, Southern Methodist University, spring 1969.

Resignations

Charles F. Austin, School of Business Administration, The American University, June 1968.

Edwin G. Caudill, School of Business Administration, The American University, July, 1968: Guilford College.

Charles A. Dailey, School of Business Administration, The American University, June 1968: Dartmouth.

Leslie Fishman, University of Colorado, June 1969: University of Warwick, Coventry, England.

Virgil Ketterling, University of North Dakota.

Robert F. Lanzillotti, Michigan State University, July, 1969.

Mortin J. Plotnik, Illlinois Wesleyan University.

Clark Puckett, State University of New York, June 1969.

John Owen Quigley, School of Business Administration, The American University, June 1968.

Miscellaneous

Leland L. Johnson, research director of the Presidential Task Force on Communications Policy: senior economist, The RAND Corporation.

The Yugoslav-American Commission in Belgrade has nominated four Yugoslav economists for travel grants to the United States. Each individual is available for a teaching or research position during the academic year, 1969-70. For detailed information, please contact Georgene B. Lovecky, Executive Assistant, Committee on International Exchange of Persons, 2101 Constitution Ave., N.W., Washington, D.C. 20418.

The Committee on International Exchange of Persons Conference Board of Associated Research Councils will issue, in March 1969, a list of foreign scholars available under the provisions of the Fulbright-Hays Act for appointments in American colleges and universities during the academic year, 1969-70. Fulbright-Hays travel grants cover costs of round-trip transportation from the home country to the United States, provided arrangements can be completed for a lecturing or a research appointment with appropriate stipend at an American institution of higher learning. Information concerning the procedures for extending invitations and the conditions of appointment may be obtained from: Miss Grace E. L. Haskins, Program Officer, Committee on International Exchange of Persons, 2101 Constitution Ave., N.W., Washington, D.C. 20418.

The University of Colorado has received a grant from the National Science Foundation for a summer institute in Computer Science in Social and Behavioral Science Research for college teachers in the social and

behavioral sciences. The institute will be held in Boulder, Colorado for five weeks, beginning June 16, 1969 and ending July 18, 1969. Each participant, a college or university teacher, will receive a stipend of $500 plus travel and dependency allowances. The faculty of the institute are: Bert F. Green, Edward A. Feigenbaum, William R. Uttal, Philip J. Stone, Claude McMillan, Daniel E. Bailey, and Richard D. Duke. Applications are to be completed on or before April 15, 1969. For application forms and more detailed information, write to: Daniel E. Bailey, Ketchum 8, University of Colorado, Boulder, Colorado 80302.

SENIOR FULBRIGHT-HAYS AWARDS IN ECONOMICS AND BUSINESS ADMINISTRATION FOR 1970–71

Applications should be filed by June 1 for Fulbright-Hays appointments for university lecturing and advanced research abroad during 1970-71. About 30 scholars in economics and business administration are serving as lecturers during 1968-69 in Argentina, Australia, Ceylon, Germany, Greece, India, Iran, Italy, Japan, Korea, Liberia, Malaysia, Mexico, Peru, Republic of China, Turkey, the United Kingdom, Uruguay, and Yugoslavia, and scholars in these fields have been granted Fulbright awards for research during 1968-69 in Australia, Germany, the Netherlands, and Spain. Appointments for 1970-71 may be fewer in number because of reductions in the appropriation for the Fulbright program.

The basic application requirements are: U.S. citizenship; a doctoral degree or equivalent status for research; college or university teaching experience for lecturing appointments; and in some cases, proficiency in a foreign language.

Senior Fulbright awards ordinarily consist of a maintenance allowance in local currency to cover normal living costs of the grantee and family while in residence abroad, and roundtrip travel for the grantee (transportation is not provided for dependents). For lecturers going to most non-European countries, the award includes a dollar supplement, subject to the availability of funds; or it carries a stipend in dollars and foreign currency, the amount depending on the assignment, the lecturer's qualifications, salary, and other factors.

For lecturing in 1970-71, application before June 1, 1969 is strongly recommended. Research applications will not be accepted after that date.

Application forms, a list of openings, and details on the terms of awards for particular countries are available from the Committee on International Exchange of Persons, 2101 Constitution Avenue, N.W., Washington, D.C. 20418.

The American Economic Review

THE AMERICAN ECONOMIC ASSOCIATION

Founded in 1885

• Published at George Banta Co., Inc., Menasha, Wisconsin.

• THE AMERICAN ECONOMIC REVIEW, including four quarterly numbers, the *Proceedings* of the annual meetings, the *Directory* and Supplements, is published by the American Economic Association and is sent to all members five times a year, in March, May, June, September and December.

• Membership dues of the Association are $10.00 a year, $8.00 of each year's dues being in payment of a year's subscription both to the *Review* and the *Journal of Economic Literature*. Subscriptions by nonmembers are $10.00 a year. Single copies of the *Review*, $2.00; of the *Journal*, $1.50.

• Correspondence relating to the *Paper and Proceedings*, the *Directory*, advertising, permission to quote, business matters, subscriptions, membership and changes of address may be sent to the secretary, Harold F. Williamson, 629 Noyes Street, Evanston, Illinois 60201. To be effective, notice of change of address must reach the secretary by the 1st of the month previous to the month of publication.

• Second-class postage paid at Menasha, Wisconsin. Printed in U.S.A.

LLOYD APPLETON METZLER

DISTINGUISHED FELLOW

1968

We honor Lloyd Metzler, a theorist's theorist, for his original, penetrating, and fundamental contributions to many branches of economic theory: general equilibrium, international trade, income and employment, public finance, and business cycles. The durability of his insights and the power of his analysis are attested by the major role his papers play in the education of new generations of economists and in the cumulative advance of economic science. This has always been Lloyd Metzler's prime goal.

DISTINGUISHED FELLOWS
(In addition to Past Presidents)

1965

Edward H. Chamberlin
Harold Hotelling

1966

Abba P. Lerner

1967

Jacob Marschak

Lloyd A. Metzler

$45 Billion of U.S. Private Investment Has Been Mislaid

By ROBERT J. GORDON*

The long term behavior of the capital-output ratio is an unsolved mystery in the annals of U.S. economic growth.[1] Although private output and capital grew at roughly the same rate, both before 1929 and after 1948, between those dates a 57 percent increase in real private output was achieved with a growth of only one percent in the private nonresidential capital stock. Thus between 1929 and 1948, which were both high utilization years with similar employment rates, the marginal capital-output ratio appears to have been only .02, compared to an average capital-output ratio in 1929 of about 2.4.[2] How did the U.S. economy succeed in producing so much extra output with so little apparent addition to its stock of plant and equipment during this 20 year period?

The question has numerous answers, and many of them do not involve subtleties in economic theory nor econometrics, but instead, glaring errors in the measurement of the private capital stock. Imperfections in previous capital estimates (including [8], [14], and [18]) are important enough to raise serious questions about the validity of all production function studies of the U.S. economy which extend before 1948, e.g. [5], [20], and [21]. This paper attempts to go beyond destructive criticism of the official figures to the presentation of new capital estimates based on primary source data.[3] The subject here is a little known $45 billion treasure chest of plant and equipment which the U.S. Government has purchased for the use of private firms since 1940, but which has never been counted as part of the private U.S. capital stock.[4] The major conclusion is that the neglected assets explain a substantial portion of the decline in the manufacturing capital-output ratio during World War II and the early postwar years, since the addition of these assets results in a 33 percent increase in the manufacturing capital stock between 1940 and 1945, as compared to the two percent increase shown in previous data.

I. *Government Financed Assets in the National Income Accounts*

Most productivity studies are rightly confined to the private sector and ignore government output, which is omitted because of its artificial treatment in the National Income and Product (*NIP*) accounts. Government output, which cannot be valued at a market price because it

* The author is assistant professor of economics at the University of Chicago and a research staff associate of the National Bureau of Economic Research. He gratefully acknowledges research support from a Ford Foundation dissertation fellowship and from the Project for Quantitative Research in Economic Development at Harvard University.

[1] Other aspects of its behavior have been discussed by Anderson [1].

[2] Capital data in 1958 prices from [14, p. 46, column 1]. Output from [6, Table B-8, p. 219].

[3] For criticisms of the official capital data, see [12] and [13]. In addition to government financed assets, the author's capital measurement study emphasizes errors in the measurement of service lifetime and thus differs from the related study by Jorgenson and Griliches [17] which ignores both of these subjects. For a critique of Griliches-Jorgenson, see [9]. Another paper in the series is [10].

[4] The figure of $45 billion is derived in Table 4 and represents an estimate of cumulative 1940–65 U.S. government expenditures on privately operated plant and equipment (in 1958 prices) minus the small portion already included in the official *OBE* capital stock data.

is not sold, is measured simply as the compensation of government employees. Thus real government output, displayed in the national accounts in 1958 dollars, is a bogus concept. Since the deflator is average employee compensation rather than an actual price, real government output grows by definition at exactly the same rate as labor input, and productivity by definition never changes. The government sector, therefore, must be excluded in any project which intends to study productivity change.

All of this is well known, of course, and most productivity studies consequently limit their analysis to private output, which is assumed to be produced with private inputs. Unfortunately, the official data [14] [26] do not define government and private investment and capital by the sector in which they are used, as required for productivity analysis, but rather by the sector which finances them. This creates no problems with regard to government output which is not directly produced by privately financed capital. But structures and equipment financed by the government have played a surprisingly important role during the last 25 years in the production of private output.

Government capital is operated by private firms under varying types of agreements. Many firms operate on a normal profit making basis, using government financed machinery in their own factory buildings. In other cases, the government owns the equipment and the building, and the firm operates the plant for the government for a fixed fee, with any profits or losses accruing to the government.[5] In addition, many plants which the government was unable to sell after World War II have been leased to private firms who operate them on a profit making basis. All of these cases, despite their differences, have a crucial similarity. The wages and salaries of employees in the government owned plants are paid by private firms, and these payments are counted, along with the fees or profits earned by the operators, as part of national income originating in the private sector. Any productivity study is misleading if it includes all of private output but neglects part of the capital which was used to produce that output. This, unfortunately, is the fate which has befallen all previous studies, because the U.S. national accounts have failed to keep track of government capital. Government purchases of industrial structures and equipment are lumped in with all other government current purchases of goods and services, and there is no published series on the amount of government capital which has been used during the past 25 years to produce private output.

The amounts involved are not trivial. In the spring of 1945, private firms were producing steel ingots at a full capacity rate, but they owned only 90 percent of their fixed capital input; the other 10 percent was owned by the government. Boeing, Grumman, Republic, and other private firms were rolling out aircraft at an unprecedented pace, but the planes were rolling out through government owned doorways and the supervisors congratulated each other from government owned chairs and desks. Nor were government facilities merely a wartime phenomenon. After the war in 1947, for example, Alcoa, Reynolds, and Kaiser produced half of the nation's aluminum output in government owned plants. In 1951, about half of our rubber supply was synthetic, and all synthetic rubber was

[5] The cost-plus-fixed-fee (*CPFF*) contract, mainly used in World War II, was considered more efficient than the cost-plus-percentage-profit scheme, which was more common in World War I. The latter arrangement, of course, gave profit-maximizing munitions makers an incentive to keep costs as high as possible. More than $50 billion in *CPFF* contracts were signed in World War II. See [15, p. 86] and [22].

produced by private firms in government owned plants. Even today, 123,000 employees of private firms work in plants and laboratories owned by the *Atomic Energy Commission*, having a gross book value of over $8 billion. Much of the aircraft and ordnance production for the Vietnam war has been carried on by private firms with government owned plant and equipment. And, similarly, almost all of the construction, testing, and firing of rockets in the U.S. space program has been performed by private contractors in government owned facilities.[6] In addition, the production of commercial aircraft is carried on partially through the use of government financed equipment.[7]

The majority of the government industrial facilities in question were constructed during World War II. To account for all of the government capital which has been used since then to produce private output, three separate questions must be answered: How much government capital was used by private firms during World War II? How much of that capital continued in operation by private firms after the war? And how much capital has been built by the government since the war for private operation? The distinction between World War II facilities and capital built after the war is dictated by the inadequacy of the available statistics. No complete record is available which traces the eventual disposition of all World War II facilities, and a large part of the work reported in this paper is an attempt to infer from scattered pieces of evidence what happened to these plants after 1945.

At present the *NIP* accounts, the official estimates of wartime and postwar capital formation, do not include government capital used by private firms but only privately financed structures and equipment purchases. Government financed facilities enter the private capital stock only when they are sold to private firms, and at the time of sale they are added to private capital formation at the sales price, not at original cost. This approach results in an inappropriate measure of the capital actually used by private firms, for several reasons:

1. The timing in the *NIP* accounts is wrong for productivity analysis. The World War II facilities were used by private firms from the date of construction, mainly in 1941–43, not from the date of sale in 1946–48. Thus the capital available for use in private production during the war has been understated by a large amount.

2. The official valuation at the time of sale is wrong. The *NIP* accounts have valued the facilities at their sale price, not at original cost. Since average sale prices after the war averaged only about 30 percent of original cost, the *NIP* estimates seriously understate the cost of the assets, which have thus been included in gross capital formation at their net depreciated value.[8] In the investment statistics they are completely merged with all new capital

[6] Among the sources for this paragraph:

Iron and Steel. Capacity was expanded from 81.6 million tons in 1940 to 95.5 million tons in 1945. See [2, p. 19]. Of total expenditures on basic iron and steel facilities expansion from 1940 to the end of 1944, the government-financed share was 66 percent [33]. Sixty-six percent of the expansion of 13.9 million tons is 9.2 million tons, or roughly 10 percent of 1945 capacity.

Aircraft. The Government financed 88 percent of the World War II expansion of aircraft plants [33].

Aluminum. [31, Second Quarter 1947, p. 8].

Rubber. Production figures from [29, April 1, 1951, p. 16].

Space. For instance, the first stage of the Saturn I and V rockets are produced by Chrysler and Boeing in a Louisiana plant originally built to produce tanks in World War II. See [24, pp. 230 ff.].

[7] For instance, regarding a particular piece of government-owned equipment, it was recently revealed that ". . . about 1 year after an 8,000 ton forge press costing $1,400,000 was installed, it was used extensively for commercial production of jet engine midspan blades. In the 3 year period ended December 31, 1965, the 8,000 ton press was used 78 percent of actual production time for commercial work . . ." [24, p. 416].

[8] For the 30 percent figure see Table 3.

built in the year of sale, e.g. 1947, and for depreciation and retirement calculations are treated as if their age was the same as new 1947 assets.

3. In addition, since used goods sold by the government in 1947 are merged with new capital built in 1947, they are deflated in 1947 prices. This is wrong, since they were built in 1941–45 and should be deflated in the prices of the year of construction.

4. The procedure makes no allowance for capital which was not sold. Many of the World War II facilities, as we shall see, were either leased to private operators or continued to be operated by private firms for the government.

5. No account is taken of the government industrial structures and equipment which were built after the war. These have become a more and more important part of the capital stock in the late 1950's and 1960's as much of the World War II equipment has worn out and has been abandoned.

Only one previous investigator has made any sort of allowance for the peculiarities of wartime facilities expansion. In his Ph.D. dissertation [16] Patrick Huntley did not make any allowance for the faulty *NIP* treatment of government capital but did take a step backwards by excluding part of wartime private investment from the postwar capital stock. During World War II, manufacturers were allowed five year amortization when they expanded facilities for war purposes. They were allowed to decide for themselves whether their investment was in a normal peacetime or extraordinary wartime pattern. Using statistics on wartime amortization, Huntley excluded fully half of 1940–45 private investment from his estimates on the grounds that "A manufacturer who elected to 'amortize' his entire plant and equipment expenditure made during the war presumably registered the opinion that his entire expenditures were for manufacturing facilities of wartime usefulness having limited peacetime use" [16, p. 72]. But a private manufacturer would have every incentive to reduce his tax payments by declaring that his facilities were for wartime purposes. Even if he expected wartime tax rates and profit rates to continue after the war, the present value of tax deductions from five year straight line tax amortization would be more than double that of 25 year straight line depreciation (at a 10 percent discount rate). Further, the manufacturer doubtless would assume that tax rates would be reduced after the war and he would try to maximize his amortization deductions during the high tax war years. In addition, he would tend to underestimate the peacetime usefulness of his newly expanded plant. Profoundly discouraged by vivid memories of Depression, most manufacturers in the early 1940's never dreamed that wartime levels of demand would continue after the war. The return of normal peacetime conditions, they thought, would only bring profitless excess capacity to their expanded 1943 plants.

There are several scattered pieces of evidence which conflict with Huntley's assumption that much private investment was for temporary wartime purposes. After the war, Deming and Stern made a careful study of the peacetime uses of plants built during the war in the Southern states and wrote that "private industry in most cases financed facilities that were closely related to normal operation" [4, p. 24]. Creamer, in his study of manufacturing capital, also concluded that privately financed assets built during World War II were used for production after the war and his net capital stock estimates for the postwar years were raised above book value figures by the amount of this excessive wartime amortization [3, pp. 219–20]. Further, an inspection of the industrial

composition of public and private wartime expansion confirms that almost all extraordinary wartime needs were financed by the Government rather than private industry. In the war-oriented explosives, ammunition, ordnance, aircraft, and ships categories, for instance, the Government financed 93 percent of facilities expansion and private firms the remainder. Over 90 percent of privately financed expansion was in the food, textiles, paper, basic chemicals, petroleum, rubber, basic metals, and machinery industries, and these expanded facilities were available to meet the surprisingly buoyant demands for civilian goods in the prosperous postwar economy.[9]

The Korean war practice of five year amortization should not mislead us into thinking that plants built during 1951–53 were discarded after five years. While amortized over five years, new capacity in steel, aluminum, and other basic industries constructed in 1951–53 to meet the Korean War emergency was just as useful as plants built during any other period in meeting the needs of the post Korean economy. In both wars, rapid amortization was a device to induce investment in industries in which there seemed a large possibility of a postwar slump in demand. Thus, contrary to Huntley's procedures, no special allowance will be made here for the premature discarding of privately financed facilities built in either war. In fact, the calculations below allow us to estimate the extent of premature retirements, and they appear to have been relatively slight even for government financed facilities (see Table 2, line 2b).

Most of the results presented below are necessarily subject to a substantial margin of error. The requisite data are scattered and difficult to locate. One government agency after another spent a brief period collecting figures, and each during its tenure changed definitions and lost track of some assets. Responsibility for data collection shifted from the War Production Board to the Civilian Production Board, the War Assets Administration, and later the Defense Department and other branches. Many of the records legalistically divide assets into real property and personal property, which are useless distinctions for productivity analysis and which, unfortunately, do not correspond to structures and equipment. When an entire plant was sold to one bidder, its installed equipment was considered real property, but if the same equipment was removed to be sold separately, it was tallied as personal property. Personal property does not consist entirely of equipment, but includes both consumer goods and intermediate products. Because of these difficulties, our breakdown into structures and equipment is only partly based on hard facts, and many of the estimates rest instead on an intricate web of guesses, extrapolations, and blow-ups based on scattered hints in obscure publications.[10]

II. *Facilities Built During World War II*

A. *Wartime Plant Expansion*

Between mid-1940 and 1945 American manufacturers spent more on new plant and equipment than they had purchased in the entire pre-war decade of the 1930's—a total of $11.4 billion (valued at original cost). While substantial, this expenditure in real terms was only 16 percent of the 1940 manufacturing capital stock, a moderate increase compared to the 95 percent rise in real manufacturing output from 1940 to 1944.[11] Such an incredible expan-

[9] Industrial composition figures are from [33, inside front cover].

[10] Because of space limitations, many tables and source notes to tables are omitted from this summary article and will appear later in book form.

[11] Actually, this 16 percent increment was not enough to offset retirements in calculations of the capital stock

TABLE 1—MANUFACTURING STRUCTURES AND EQUIPMENT BUILT BETWEEN JULY 1, 1940 AND DECEMBER 31, 1945

($ Million, Original Cost)

	Financed By	Operated By	Total	Structures	Equipment
1.	Private	Private	11,435	3,162	8,273
2.	Government	Private	15,985	7,665	8,320
3.	Government	Government	1,656	1,030	626
4.	Total		29,076	11,857	17,219

Sources By Line:

1. [27, pp. 48, 100].
2. Structures: Public industrial construction [26, Table 5.4, line 33], adjusted for government-operated plants.
Equipment: [32, pp. 13–16], adjusted to include Project Manhattan expenditures and to exclude expenditures for government-operated plants.
3. Data reported for 1955 [23, p. 72] and adjusted for estimated 1945–55 expenditures and retirements, and for estimated construction of atomic energy facilities operated by the Government.

sion of production with so little new capacity, long considered the epochal American *Wirtschaftswunder*, is less of a miracle than first appears. For conventional statistics on private investment by manufacturers completely ignore almost 60 percent of the wartime expansion of privately operated facilities. The forgotten billions of plant and equipment purchased by the government for private plant operators are shown in line 2 of Table 1. Some of the government financed facilities were operated by the government itself and are of no interest in this paper, for their output is not included in the private *GNP*. Workers in government operated arsenals and shipyards are on government payrolls, and their wages are part of national income originating in the government sector. Line 3 of Table 1 shows these assets separately.

It is not easy to find out what happened after the war to all of the assets in line 2 of Table 1. How many of them were specially built for war purposes and were tossed onto the scrap heap during the post V-J day reconversion, and what proportion remained in the capital stock to produce private output after the war? One method of tracing the disposition of these facilities is obvious; a complete list of the plants and their original cost is available, and a man with lots of time and pockets stuffed with airline tickets could set out across the country to find out what has happened to them since 1945. He would find the plants in a variety of uses: some may be producing the same product as in wartime, some may have been transferred to operators in completely different industries or to public agencies, and others may have been subdivided among a number of users.[12] While most of the structures could probably be tracked down in this way,

based on constant lifetimes, and *OBE* estimates of the manufacturing capital stock show a decline from 1941 to 1944. In truth, of course, many retirements may have been delayed during these years, and the actual capital stock (based on unrevised investment data) probably rose.

[12] Consider the varied fates of several aircraft plants. The Lockheed plant in Burbank (original cost: $46 million) was partly sold to Lockheed and partly to smaller companies in other industries. The Consolidated-Vultee plant in San Diego (original cost $46 million) was used in 1948 by several private companies, the County of San Diego, and the San Diego baseball club. In the 1950's, however, the plant was again producing aircraft. The Boeing plant in Wichita (original cost: $32 million) is still retained in the ownership of the Air Force, even though Boeing currently uses it for commercial aircraft production.

TABLE 2—GOVERNMENT FINANCED ASSETS OPERATED BY THE PRIVATE SECTOR, BY TYPE

($ Million, Original Cost)

	Total	Structures	Equipment
1. Built 1940–45 for private operators; used postwar by private operators	12,323	7,165	5,158
a. Plants sold to private firms, 1945–54	3,516	2,557	959
b. Equipment sold separately from plant	1,712	—	1,712
c. Synthetic rubber plants	488	424	64
d. Retained by defense department for private use	5,285	3,582	1,703
e. Atomic energy facilities	1,322	602	720
2. Built 1940–45 for private operators; not used postwar in private sector	3,662	500	3,162
a. Equipment sold to public sector or abroad	735	—	735
b. Residual, assumed scrapped	2,927	500	2,427
3. Used 1940–45 by military; sold postwar to private sector	1,654	—	1,654
4. Built after 1945 for private operators			
a. Defense industries	4,784	850	3,934
b. Atomic energy	5,668	1,997	3,671
c. Space program	1,717	688	1,028
5. Total	32,363	12,228	20,135

Sources: See text.

nothing could be discovered about the disposition of the equipment, much of which was removed from government financed plants and sold after the war to private users (resulting among other things in seven lean years for the U.S. machine tool industry between 1944 and 1951). Also, of course, a great deal of the unsold machinery has by now become obsolete and has been scrapped. Thus our man in the field would be able to obtain information only about structures, and the disposition of government financed equipment would have to be estimated by indirect methods. Time and money are important constraints, of course, and we shall have to leave traveling expeditions to others. Thus we are forced to employ indirect methods for estimating the postwar disposition of both structures and equipment. The following section briefly describes the methodology of the investigation, but full details are omitted to conserve space and will be published later.

B. *The Postwar Disposition of Assets Built During World War II*

The results of the study are summarized in parts 1 and 2 of Table 2, which shows that it has been possible to identify the postwar disposition of about 83 percent of the government financed facilities built during World War II. The remainder, the residual of $2,927 million listed on line 2b, represents facilities which were privately operated during the war but which could not be located after 1945. Most of the residual probably represents equipment which was specially designed for

TABLE 3—POSTWAR DISPOSAL OF GOVERNMENT PROPERTY BY THE WAR ASSETS ADMINISTRATION THROUGH JUNE 30, 1949

($ Million, Original Cost)

Category	Acquisitions	Sales	Leases	Miscellaneous Disposals	Property Available For Disposal	Sales Realization	Sales Realization as Percent of Original Cost of Property Sold
Consumer and producer goods	9,719	8,242	—	1,477	—	2,443	29.7
Real property	7,786	3,771	649	2,630	737	1,211	32.1
Aircraft and components	7,858	1,798	23	5,643	395	187	10.4
Other	1,835	1,241	—	569	24	304	20.4
Total	27,198	15,052	672	10,319	1,156	4,145	27.5

Source: [31, Second Quarter 1949, Tables 1–6, pp. 41–46].

wartime weapons which were never produced again.

1. *Postwar Sales of Real Property.* On V-J day, the government owned billions of dollars of plants, equipment, and an almost infinite variety of other goods which it no longer needed. Warehouses bulged with trucks and tanks, trousers and toothpicks. More than $27 billion of goods, equal in value to about 15 percent of the 1946 *GNP*, were declared surplus by the *War Assets Administration* (*WAA*), which had been created in 1945 and charged with an immense sales job. At the peak of its activity, the *WAA* had almost 60,000 employees busy trying to attract buyers for items in its storehouse.

The composition of *WAA* acquisitions is shown in Table 3, which is drawn from the last report published by the *WAA* before it closed up shop and disappeared. The Consumer and Producer Goods category includes consumer, intermediate, and producer goods. While some are materials such as wood and steel and are not part of fixed capital, a substantial amount consists of machinery and equipment which had been removed from government financed plants, as well as producers durables (e.g., trucks) which had been operated during the war by the Army and Navy. Real property includes structures and all equipment therein disposed of with the plant. Equipment sold to a separate bidder is classified in the Consumer and Producer Goods category. Most of the Aircraft and Components category, of course, was made up of combat aircraft which were scrapped (hence the large entry in line 3 of Table 3 under "miscellaneous disposals"), but many planes were sold to private firms as airline transports, corporate planes, and training planes in private flying schools, and hence should be included in the postwar private capital stock.

By June 30, 1949, as shown in Table 3, the *WAA* had completed most of its job. More than half of its acquisitions, about $15 billion worth, had been sold for 27.5 per cent of original cost. The relatively high sales realization for long lived real property was to be expected, as was the extremely poor return on sales of evanescent aircraft. The *WAA*'s job was made easier by the ease with which free miscel-

laneous disposals out the back door of the store reduced the size of the inventory which had to be sold to paying customers through the front door. In addition to junked aircraft, other miscellaneous disposals went to public agencies; army hospitals became state hospitals and school buildings, camps and reservations were transferred to school districts, and hundreds of army airfields were donated to municipalities, making possible a vast increase in the number of cities receiving commercial air service.[13]

There are two main obstacles to an estimate of the war-built industrial facilities which were sold by the *WAA*. First, the official tabulations, as shown in Table 3, do not subdivide real property into its industrial and nonindustrial components. A substantial part of the real property acquisitions do not represent the industrial plants we are searching for, but rather hospitals, barracks, airfields, and training camps. The second difficulty is that not all industrial property was sold to private firms; some was sold to local governments for nonindustrial purposes. The $3,156 million estimate of plant sales in line 1a of Table 2 is based on scattered hints in *WAA* reports. Of the total acquisitions of real property shown in Table 3, about 56 percent (or $4,356 million) consisted of industrial plants. Some of these were disposed of to public agencies, but the majority were sold to private firms at various dates between 1945 and 1954. The division in line 1a between structures and equipment was derived as a residual at the end of the estimation process.

2. *Equipment Sold Separately*. The figure in line 1b of Table 2 is based on an estimate of the fraction of sales of Consumer and Producer Goods (as listed in Table 3) which consisted of machinery. The major source of information was a detailed breakdown of early *WAA* acquisitions by three-digit industry. Another helpful source was a periodic *WAA* report on the disposition of machine tools. Sales to public agencies and abroad are listed separately in line 2a. A similar procedure was used to derive the estimates in line 3 of military equipment used after the war by private firms (e.g., trucks, air conditioners, airplanes, office equipment, etc.).

3. *Synthetic Rubber Plants*. At the close of the war, the government owned 44 synthetic rubber plants which had cost $670 million to build during the preceding four years.[14] These plants contributed almost 100 percent of the U.S. rubber supply after the Japanese conquest of Malaya in 1942. Disposal of the plants was delayed for many years after the war because of uncertainty about the future price of natural rubber. It was difficult to determine how much of the capacity should be kept in active operation when it was not known whether synthetic rubber could be sold permanently at a profit.

Eventually, the doubt about the future of synthetic rubber was resolved in 1950 when natural rubber prices spiralled to many times the cost of the synthetic product. Synthetic production mounted rapidly and by 1951 amounted to over 50 per cent of the total U.S. rubber supply.[15] Finally, in 1955, almost all of the remaining capacity was sold. Most of the sales were made to the same firms which had operated the plants since they were first built during World War II. These firms had operated the plants without interruption for ten years or more, since even at its low point in 1949 capacity utilization in synthetic rubber was about 50 percent,

[13] Because of airport transfers, the bargain basement sale prices of used air transports, and the release from the armed services of thousands of pilots who had been trained at government expense, much of the postwar private air transport industry was established practically without cost to the private sector.

[14] See [2, pp. 24–26].

[15] See [29].

and government policy was to operate all plants at half-throttle rather than to close down any facilities.[16]

4. *The Departmental Industrial Reserve.* Lines 1a, 1b, 1c, and 2a in Table 2 summarize the results of our study of the *WAA* data. They are quite surprising, for postwar sales account for only about $6.5 billion, less than half of the wartime government financed, privately operated industrial facilities expansion of $16.0 billion. Where are all of the missing plants? There is no mention in any of the *WAA* progress reports of the billions of dollars worth of plants which were built during the war but were never acquired by the *WAA*.

After a long search, some evidence on the missing plants has been found, allowing a crude estimate of their value and date of construction. The study reveals an enormous industrial empire which most economists have never heard of. The industrial plants owned by the Department of Defense, sometimes called the *Departmental Industrial Reserve* (*DIR*), had a 1960 book value of more than $11 billion of structures and equipment. Since many of the plants were constructed decades ago, this original cost valuation is a considerable understatement of replacement cost. The *DIR* includes many plants operated by private firms in addition to the government operated arsenal system and Navy industrial facilities. Most of the plants were built during World War II, and the facilities constructed before 1939 are mainly old government operated arsenals (most of which were established between 1777 and 1863). The plants are retained in government ownership to assure supplies of essential weapons and explosives in wartime without any lag for the construction of new plants or the conversion of civilian factories, as was necessary during the early days of World War II. About 80 percent of the value of the *DIR* is accounted for by four industries, explosives and ammunition, weapons, aircraft, and shipyards. The aircraft industry in particular, is a virtual ward of the government, having contributed approximately 10 percent of the funds for its own expansion in World War II and only about one-third during the Korean War.[17]

The *DIR* showed its worth in the Korean war, when in 1950–51 defense orders for ordnance and weapons suddenly inundated the economy. American manufacturing output grew by 34 percent between 1948 and 1953, an unusually rapid rate of growth considering that the earlier year was not a recession year but was characterized by shortages and full capacity operation. Part of the growth in capacity, of course, was due to new construction between 1948 and 1953, but a substantial amount was due to the reactivation of *DIR* plants, many of which had rested dormant since World War II. Reactivated *DIR* plants, for instance, added 10 percent to aluminum capacity in 1951, contributed substantially to increased nitrogen production, and counted for almost all of the Korean war increase in magnesium production.[18] The number of aircraft plants in operation doubled between 1950 and 1953 as old *DIR* plants came into the active list again. "Such famous World War II aircraft plants as those at Marietta, Ga., Tulsa, Okla., and Kansas City are again in production."[19] By 1952, 92 percent of the *DIR* plants were in operation, as opposed to an approximate 60 percent in 1950.[20]

The *DIR* plants have remained un-

[16] Production figures from [29, April 1, 1951, p. 16]. Government policy discussed in [30, p. 3]. Estimates in Table 2 based on [19].

[17] The general description in these paragraphs is based on [23, pp. 61–62].

[18] [29, January 1, 1952, pp. 15–17].

[19] [29, January 1, 1953, p. 27].

[20] [28, 1952, p. 40].

known because of the incredibly primitive statistical reporting standards of the Department of Defense. In most cases (e.g., in the utility and transportation industries) government regulation increases statistical information available to the public, but government ownership is often an excuse for secrecy. In the case of the *DIR*, practically no solid information is available. Congress has only recently discovered the existence of the *DIR*, and in 1967 Senator Proxmire held hearings on the private usage of government owned equipment. The hearings [24] revealed that the Department of Defense had maintained scanty records, had little idea how many of the machines were being used by private contractors for commercial work, and was receiving inadequate rental for the machines in commercial production.

The estimates in line 1d of Table 2 are based partly on these hearings, and in addition, on Annual Reports of the Department of Defense and a 1955 Hoover Commission Report [23].

5. *Atomic Energy.* The information on line 1e is based on the official history of the *Atomic Energy Commission* and can be considered relatively reliable. Most of the $1.3 billion total consists of production plants which were built by the *AEC* during the frantic 1943–45 rush to develop a working atomic bomb, and which were operated by private contractors, e.g., Union Carbide at Oak Ridge, Tennessee.

III. *Industrial Facilities Built After 1945 for Private Operation*

1. *Defense Industries.* Estimates for Department of Defense facilities constructed after 1945 were estimated simultaneously in lines 1d and 4a (see above). The post–1945 expansion of privately operated Department of Defense facilities appears to have been concentrated in equipment; the construction of new structures was limited and appears to have been largely for government operated facilities.

2. *Atomic Energy.* The *AEC* financed a large program of postwar construction, the bulk of it in the years 1951–55. The huge production complexes originally built during World War II at Oak Ridge and Hanford were joined by many new facilities, including three which cost more than $750 million apiece. Production facilities have not been expanded much since 1956, but recently there has been a substantial amount of construction of research facilities, many of which are operated by private institutions (e.g., Harvard University).

Since 1950 the *AEC* has published an annual financial report which, despite its deficiencies, puts the Defense Department to shame. Each year figures are reported for plant and equipment expenditures and the book value of production plants, research labs, and other facilities. In recent years the reports have included a complete list of all *AEC* facilities, showing for each the book value of completed projects and the cost of work under construction. While the *AEC* information is more adequate than most in this paper, there are nevertheless several deficiencies which force us to make some guesses. The years between 1945 and 1950 are a statistical dark age on which practically no information exists. In addition, none of the *AEC* reports makes any distinction between structures and equipment; all of our estimates which divide *AEC* expenditures between structures and equipment are based on a single inventory taken in 1960.

3. *Space Program.* Of the 420,000 peak employment of the U.S. space program in 1966, only 7 percent actually worked for the *National Aeronautics and Space Administration* (*NASA*).[21] The vast majority, from those building the Apollo capsule in

[21] See [25, p. 109].

Downey, California, to the men who helped the astronauts into their spacecraft at Cape Kennedy, Florida, were employees of private firms. But their tools and many of the buildings in which they worked were owned by *NASA*, which spent about $2.5 billion on facilities between 1959 and 1967.[22] The figure on line 4c of Table 2 is smaller than this, excluding facilities (like the Manned Spacecraft Center in Houston, Texas) built for use by *NASA* employees, as well as plants used by *NASA* but built by the Government in World War II for defense purposes (to avoid double counting).

IV. *Conclusion*

A. *Evaluation of the Results*

All of this adds up to over $32 billion of expenditures valued at original cost, representing a substantial stock of facilities, most of which has helped to produce private output but has never before been included among the factors of production in the private sector. The existence of this vast amount of previously unmeasured capital explains in part how the private American economy produced so much during the war and early postwar years with such a small measured increase in the stock of capital relative to the level of the late 1920's.[23] The implications apply mainly to the manufacturing sector, since about 90 percent of the assets listed in Table 2 constitute manufacturing structures and equipment.

In the research work for this paper the facilities shown in Table 2 have been separately attributed to the years in which they entered private production. This enables us to convert the values from original cost to 1958 prices, with results shown in Table 4.[24] These data are then used in the top frame of Figure 1 to illustrate changes in the real gross stock of government financed assets from 1940 to 1965. For comparison, the bottom frame shows the small portion of assets originally financed by the Government which the *OBE* has already included in its capital stock estimates.[25]

The figure indicates that the stock of government financed, privately operated assets expanded rapidly to a 1945 peak of $31 billion (in 1958 prices), fell back to about $26 billion after the war as some special purpose equipment was retired, and then expanded to its all time peak of $32.5 billion in 1954 after the Korean expansion of Defense Department equipment and *AEC* facilities. Retirements of equipment built during World War II resulted in shrinkage of the stock from 1956 to 1961, but after that date, retirements were offset by further purchases of manufacturing equipment by the Defense Department and new facilities for the space program.

The reliability of these final results is not easy to evaluate. Since all of the available evidence has been used to construct the basic estimates, nothing remains to serve as a cross check. The figures must be regarded as tentative, since they are subject to changes and improvements as further data discoveries are made. The *AEC* estimates are probably the most reliable,

[22] See [25, p. 100].

[23] While this paper presents the first estimate of the portion of government financed capital which has contributed to private production, previous estimates have been made of the original cost value of total government military and civilian capital (including military weapons and making no distinction as to the assets used in the private sector). See Goldsmith's estimates [7, pp. 88–91]. Goldsmith, however, does not measure military industrial structures, which represent about one quarter of the total value listed above in Table 2 [7, p. 395, notes to Table B-166, column 5].

[24] The equipment deflators are from the *NIP* accounts, but the structures deflators are from [10, Appendix Table A-1, column 2].

[25] The top frame of Figure 1 was derived from Table 4. The bottom frame was derived from [26, Table 5.3, line 28 and Table 5.5, line 4]. In the bottom frame, structures are assumed to have a 40-year service lifetime and equipment, a 17-year service lifetime.

TABLE 4—EXPENDITURES ON, AND RETIREMENTS OF, GOVERNMENT FINANCED STRUCTURES AND EQUIPMENT, BY YEAR OF ENTRANCE INTO AND EXIT FROM PRIVATE OPERATIONS, 1940–65

($1958 Millions)

	Expenditures				Retirements			
	Manufacturing		Nonmanufacturing		Manufacturing		Nonmanufacturing	
Year	Structures	Equipment	Structures	Equipment	Structures	Equipment	Structures	Equipment
1940	298	228	0	0	0	0	0	0
1941	2,680	1,708	0	0	0	0	0	0
1942	5,919	5,361	2	2	0	0	0	0
1943	2,937	5,012	27	33	0	0	0	0
1944	1,705	2,335	64	91	0	0	0	0
1945	1,108	1,569	44	327	0	0	0	0
1946	13	209	21	1,290	500	6,436	0	18
1947	7	164	18	1,051	0	112	0	18
1948	14	125	25	389	0	90	0	18
1949	16	109	29	77	0	112	0	18
1950	21	128	34	63	0	62	0	20
1951	18	402	32	64	0	88	0	34
1952	198	1,235	68	113	0	248	0	32
1953	383	1,480	62	102	0	248	0	26
1954	622	1,245	34	61	95	626	0	24
1955	548	895	36	57	95	616	0	420
1956	165	418	3	17	95	212	0	426
1957	57	273	12	28	90	620	0	428
1958	40	271	48	89	225	662	0	430
1959	0	134	69	89	225	910	0	464
1960	0	149	56	95	225	902	0	470
1961	49	233	75	133	225	841	0	466
1962	252	607	89	118	225	788	0	58
1963	224	435	110	139	150	1,254	0	18
1964	465	1,110	147	172	200	608	0	2
1965	523	252	103	130	150	608	0	2
Cumulative Totals	18,262	26,087	1,208	4,730	2,500	16,048	0	3,392

Total Expenditures	50,287
minus: Amount already included in national accounts	− 4,800
Net addition	45,487

Source Notes:

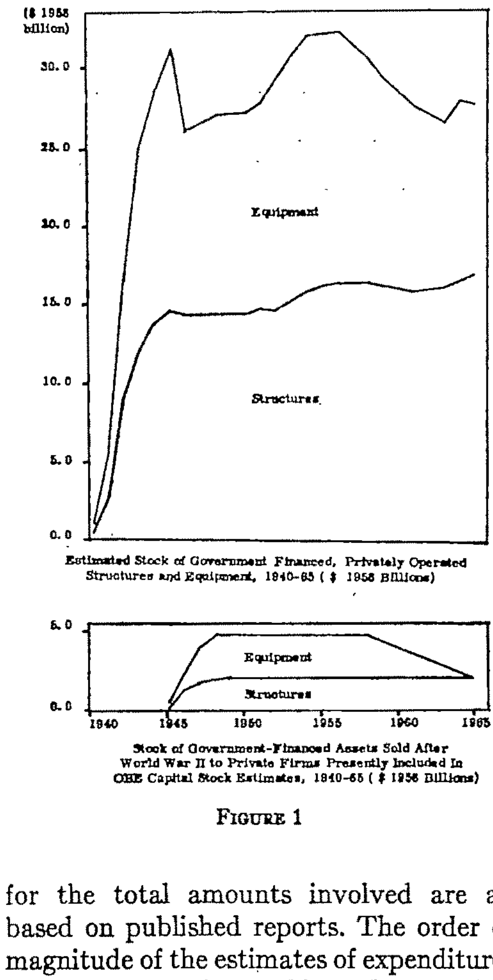
1940–45 expenditure estimates for equipment were based on [32, pp. 13–16], and for structures on [26, Table 5.4, line 33]. Post-1945 expenditure estimates were inferred from [24, p. 382] and from annual reports of the Department of Defense and the AEC. Nonmanufacturing expenditures represent primarily atomic energy laboratories and producer's durables sold to private firms immediately after World War II.

Annual retirement figures were calculated in three steps:

1. For assets which remained in the possession of the government after World War II, preliminary estimates of retirements were based on cumulated investment expenditures valued at original cost minus the appropriate original-cost capital stocks. The stock estimates were derived from the list of sources in the preceding paragraph and from [23].

2. For manufacturing structures half of the preliminary retirement estimates were assumed to have represented sales to private firms, and this portion therefore was excluded from the final retirement estimate.

3. For assets which were originally financed by the government but sold to private firms after World War II, the following arbitrary retirement patterns were assumed: no structures retirements during the period of study, on the assumption of a 40-year service life; manufacturing equipment retirements spread evenly between 1957 and 1965 (assuming an approximate 17-year life) and nonmanufacturing equipment retirements spread evenly between 1955 and 1961 (about a 14-year life).



Estimated Stock of Government Financed, Privately Operated Structures and Equipment, 1940–65 ($ 1958 Billions)

Stock of Government-Financed Assets Sold After World War II to Private Firms Presently Included In OBE Capital Stock Estimates, 1940–65 ($ 1958 Billions)

FIGURE 1

for the total amounts involved are all based on published reports. The order of magnitude of the estimates of expenditures on *DIR* plants is roughly right, and important errors in the estimation procedure are confined to the allocation of construction expenditures between the World War II and Korean years. The inclusion of some of the *DIR* plants producing ammunition and weapons gives the postwar capital stock an upward bias during 1946–50 when Defense Department procurement was very low. Purchases were at a high level throughout the 1950's and 1960's, reducing the bias substantially and eliminating it completely during the Korean and Vietnam wars.

Another possible bias, much harder to assess, may be caused by the valuation of all assets at their original cost. The returns received by the *WAA* on plants sold after the war were bound to be below original cost because of depreciation, of course, but a more important cause of the low realizations (32 percent for real property) was undoubtedly a general lack of certainty that wartime prosperity could be maintained. Businessmen were hesitant to commit large sums on war built plants which they feared might be badly underutilized as peace broke out.[26] Thus, many firms acquired modern, well-equipped plants in 1946 at bargain prices which seriously understate their full capacity ability to produce. Part of this plant capacity may have gone unutilized at first, but utilization doubtless improved in later years as private demand and output increased far above 1946 expectations.

Any upward bias in our estimates for these reasons could not be large relative to the total amounts listed in Table 2, since it would only apply to the $2.6 billion of structures sold after the war. Pieces of equipment, unlike structures, are movable, and many were taken from war production plants and sold to other users. These machines were easy to substitute for new machines built immediately after the war, and the machines purchased from the government took away the markets of the machine tool industry, leaving it in a depressed condition until 1951. Further, any such upward bias in the estimates may be partially or completely offset by significant underreporting of the basic magnitudes in government property inventories.[27]

[26] The lack of confidence during the early postwar years was also evidenced by the large gap between the high dividend yields on equities and the low interest rates on long term bonds.

[27] Consider the following example of underestimation in the government inventories: "In some instances government-owned tooling was not identifiable by physical markings or in the property records, as prescribed by the Armed Services Procurement Regulation. At one plant, government-owned tooling acquired under supply contracts at an estimated cost of $55,000,000, starting in 1952, was not controlled under a system of

As stated in the introduction to this paper, the existence of government-financed but privately operated facilities has important implications for the analysis of changes in total factor productivity and the capital–output ratio. Table 5 illustrates the effect on the U. S. manufacturing capital stock and capital–output ratio of the inclusion of the government-financed assets described in this paper, and of the new structures deflators recently derived elsewhere [10]. First, line 1a illustrates the rate of growth of the manufacturing capital stock using the unrevised data of the U.S. *Office of Business Economics* (*OBE*). Line 1b shows the growth in the stock computed with revised structures deflators, and the effect of this changeover is calculated in line 2a. Similarly, line 1c amends line 1b by adding to private manufacturing capital the government assets estimated in this paper, and line 2b shows the effect of this revision. The new deflator raises the rate of growth of capital during all three subperiods, with an average increase of .60 percent for the entire 1929–65 period. The inclusion of government–financed capital contributes a smaller revision for the entire period, .28 percent, but significantly raises the rate of growth of capital in the earliest subperiod relative to the postwar subperiods. In the unrevised *OBE* data (line 1a), for instance, capital growth from 1929–48 was only one-fourth as rapid as in the recent 1955–65 subperiod, but according to the revised data (line 1c) was four-fifths as rapid. Similarly, according to the unrevised *OBE* data, the capital–output ratio (line 4a) declined between 1929 and 1948 by 165 percent more than between 1955 and 1965, whereas the revised data (line 4b) indicate a 1929–48 decline which was only 36 percent more than in the latest period.

monetary control accounts, had never been inventoried, and lacked proper identification in the stock records" [24, p. 417].

The figures in Table 5, then, suggest that the decline in the manufacturing capital–output ratio between 1929 and 1948 was not radically faster than during recent years, but was at a rate of the same order of magnitude, about one percent per year. While suggestive, these results should not be regarded as final. The revisions suggested in this paper and in [10] correct just two of the numerous defects of the *OBE* capital data, and detailed new capital stock estimates will not be published until further research has been completed. In particular, all previous perpetual inventory capital estimates have assumed that the service lifetimes of fixed assets have been constant over the past one hundred years, and the next stage of our capital measurement study will explore the effects of relaxing this assumption.

B. *Suggestions for Future Research*

The existence of a substantial stock of government financed assets used by private firms, particularly in manufacturing, suggests a number of interesting topics for further research. For instance, the rate of return on corporate assets in manufacturing during and immediately after World War II was considerably higher than in the 1950's and 1960's. While this was partly due to high utilization rates and the inflation of sales prices relative to asset values, a further factor may have been the substantial profits earned with government financed assets which did not enter into the capital base on which the rate of return was calculated. This phenomenon may have been especially important in the aircraft, shipbuilding, and ordnance industries. In an extreme case, a company operating only government financed facilities on a cost-plus-fixed-fee basis could have achieved a virtually infinite rate of return on its privately supplied capital.

Government financed assets also present interesting research tasks for experts in

TABLE 5—EFFECT OF REVISIONS ON THE GROWTH OF ALTERNATIVE GROSS PRIVATE CAPITAL STOCKS IN U. S. MANUFACTURING, SELECTED INTERVALS

(Annual Average Percentage Rates of Growth, With All Series in 1958 Prices)

	1929–48	1948–55	1955–65	1929–65
1. Alternative Capital Stocks				
a. Old structures deflator, government-financed capital excluded	.67	3.38	2.50	1.70
b. New structures deflator, government-financed capital excluded	1.27	4.18	2.95	2.30
c. New structures deflator, government financed capital included	2.07	4.07	2.56	2.58
2. Effect of Revisions on Growth Rate of Capital				
a. New structures deflator	.60	.80	.45	.60
b. Inclusion of government-financed capital	.80	−.11	−.39	.28
3. Gross Manufacturing Output	3.35	4.80	3.51	3.67
4. Capital-Output Ratios				
a. Old deflator, government capital excluded	−2.68	−1.42	−1.01	−1.97
b. New deflator, government capital included	−1.28	− .73	− .95	−1.09

Sources by Line:

1a. Gross capital stocks were calculated by the perpetual inventory method, using investment data for manufacturing from [27], and using service lifetimes of 40 years for structures and 17 years for equipment, with a "one-horse shay" retirement distribution. Investment deflators are those in [27], which are the same as the structures and equipment deflators used to deflate investment in the regularly published U. S. National Accounts (see [6], p. 212).

1b. Same as 1a, with official structures deflator replaced by the author's Final Price of Structures Index [10, pp. 427–8, column 2].

1c. Same as 1b, with the addition of manufacturing expenditures and subtraction of manufacturing retirements from Table 4, and the subtraction of the amounts already included in the *OBE* Capital Goods study, as shown in the bottom frame of Figure 1.

2a. Line 1b minus line 1a.

2b. Line 1c minus line 1b.

3. 1929–47 from [18, Table D-II, pp. 465–6], linked in 1947 to values from [11, Table 2, p. 23].

4a. Line 3 minus line 1a.

4b. Line 3 minus line 1c.

industrial organization. Do firms which win defense contracts obtain an unfair advantage over their competitors when they receive government purchased machinery, usable not just in defense production, but also in commercial work? And is there an unwarranted concentration of government defense contracts and attendant government owned equipment in very large firms?

Finally, this study suggests that the Federal Government should attach some urgency to the task of constructing government investment and capital accounts. The demand by economists for government capital data has not been very high in the past, because the absence of meaningful government output figures has precluded productivity studies of the government sector. But the present article sug-

gests that many government assets are important inputs in the private sector.[28] In addition to the assets discussed above, for instance, much of the apparent decline in the U. S. private capital–output ratio may have been due to the gradual replacement of privately owned railroad assets by government–owned highways and airports.

[28] Government enterprises are another link between government capital and gross private product, since their output is included in the private sector, but their expenditures on structures and equipment are counted as current government spending. For this reason, national income originating in government enterprises should be excluded from private output in any production function study of the U.S. economy.

REFERENCES

1. P. S. ANDERSON, "The Apparent Decline in Capital-Output Ratios," *Quart. Jour. Econ.*, Nov. 1961, *75*, 615–84.
2. J. A. COOK, *The Marketing of Surplus War Property.* Washington 1948.
3. D. CREAMER, S. P. DOBROVOLSKY, AND I. BORENSTEIN, *Capital in Manufacturing and Mining: Its Formation and Financing.* Princeton 1960.
4. F. L. DEMING AND W. A. STERN, *Disposal of Southern War Plants.* Mimeo. MS, National Planning Association, 1949.
5. P. A. DAVID AND T. VAN DE KLUNDERT, "Biased Efficiency Growth and Capital-Labor Substitution in the U.S., 1899–1960," *Am. Econ. Rev.*, June 1965, *55*, 357–94.
6. *Economic Report of the President.* Washington 1968.
7. R. W. GOLDSMITH, *The National Wealth of the United States in the Postwar Period.* Princeton 1962.
8. ———, *A Study of Saving in the United States.* Princeton 1955.
9. R. J. GORDON, "The Disappearance of Productivity Change," Harvard Institute of Economic Research, Discussion Paper 44, Sept. 1968.
10. ———, "A New View of Real Investment in Structures," *Rev. Econ. Stat.*, Nov. 1968, *50*, 417–28.
11. J. J. GOTTSEGEN, "Revised Estimates of GNP by Major Industries," *Surv. Curr. Bus.*, April 1967, *47*, 18–24.
12. Z. GRILICHES, "Capital Stock in Investment Functions: Some Problems of Concept and Measurement," in F. Christ *et al.*, *Measurement in Economics*, Stanford 1963, pp. 115–37.
13. ———, "Notes on the Measurement of Price and Quality Changes," in *Models of Income Determination*, Nat. Bur. Econ. Research Stud. in Income and Wealth Vol. 28, Princeton 1964, pp. 281–318.
14. L. GROSE, ROTTENBERG AND R. C. WASSON, "New Estimates of Fixed Business Capital in the United States, 1925–65," *Surv. Curr. Bus.*, Dec. 1966, *46*, 34–40. Further data listed in *Surv. Curr. Bus.*, Feb. 1967, *47*, 20–4.
15. R. G. HEWLETT AND O. E. ANDERSON, *The New World, 1939/1946.* University Park 1962.
16. P. R. HUNTLEY, "State Distribution of Manufacturers' Plant and Equipment in Place, 1954–56." Unpublished dissertation, Univ. North Carolina, 1960.
17. D. W. JORGENSON AND Z. GRILICHES, "The Explanation of Productivity Change," *Rev. Econ. Stud.*, July 1967, *34*, 249–84.
18. J. W. KENDRICK, *Productivity Trends in the United States.* Princeton 1961.
19. T. E. SALISBURY, "The Disposal of Government Surplus Property: A Comparative Study—Aluminum and Synthetic Rubber." Unpublished Masters dissertation, M.I.T., 1956.
20. R. M. SOLOW, "Technical Change and the Aggregate Production Function," *Rev. Econ. Stat.*, Aug. 1957, *39*, 312–20.
21. L. C. THUROW AND L. D. TAYLOR, "The Interaction Between the Actual and the Potential Rates of Growth," *Rev. Econ. Stat.*, Nov. 1966, *48*, 351–60.
22. R. A. TYBOUT, *Government Contracting in Atomic Energy.* Ann Arbor, 1956.
23. U.S. COMMISSION ON ORGANIZATION OF THE EXECUTIVE BRANCH OF THE GOVERNMENT, *Task Force Report on Real Property Management.* Washington, 1955.
24. U.S. CONGRESS, JOINT ECOMONIC COMMITTEE, *Economy in Government Procure-*

ment and Property Management, Hearings before the Subcommittee on Economy in Government, 90th Cong., 1st sess., Washington 1968.

25. U.S. CONGRESS, SENATE, *NASA Authorizations for Fiscal Year 1968*, Hearings before the Committee on Aeronautical and Space Science, Part I, 90th Cong., 1st sess., Washington 1967.
26. U.S. DEPARTMENT OF COMMERCE, Office of Business Economics, *National Income and Product Accounts of the United States, 1929–1965*. Washington 1966.
27. ———, "1966 Capital Stock Study." Unpublished data on investment expenditures, Washington 1966.
28. U.S. DEPARTMENT OF DEFENSE, *Annual Report*. Washington 1949–65.
29. U.S. DIRECTOR OF DEFENSE MOBILIZATION, *Quarterly Report to the President*. Washington 1951–53.
30. U.S. WAR ASSETS ADMINISTRATION, *Synthetic Rubber Plants and Facilities*. Washington 1946.
31. ———, *Quarterly Progress Report*. Washington 1946–49.
32. U.S. WAR PRODUCTION BOARD, *Summary of Construction and Plant Expansion, No. 50*. Washington 1945.
33. ———, "War Manufacturing Facilities Authorized through October, 1944," Mimeo. 2 vols., Washington 1945.

The Cost of the Draft and the Cost of Ending the Draft

By ANTHONY C. FISHER*

Popular concern over our system of military manpower procurement[1] has been stimulated recently by issues surrounding the war in Vietnam. Heavy draft calls associated with the war have provoked studies by the executive branch [12], Congress [11] and the mass media, and opposition to the system from, among others, young men subject to call.

The current controversy would seem to stem from a contradiction in the system. In law, every able-bodied male is obligated to serve in some capacity, at some time. For a variety of reasons, not all do. Since college students and those in some of the professions tend to be deferred, in some cases to the point of virtual exemption from military service, it has been alleged that the system discriminates against the low-education, low-income, underprivileged members of society. Accordingly, proposals for reform have focussed on distributing the service burden in some sense more equitably. The National Advisory Commission [12] suggests a lottery, with names drawn at random. Many people, including former Secretary of Defense McNamara, have suggested some sort of national service program under which every young person would be requested to serve in some capacity, not necessarily military. A system of equivalent service might be worked out to include such programs as the Peace Corps and Vista as well as the Armed Forces. National service proposals differ in degree of voluntarism envisioned, though most seem to call for compulsory service in some form left to the individual to decide.

Some economists, and others, have suggested a voluntary system of another sort —the price system. That is, eliminate the service obligation and let the military services compete in the labor market for volunteers.

In comparing a price system to a draft system, it is important to distinguish between money cost to the Department of Defense and the general taxpayer on the one hand, and real cost to society on the other. In a draft system, part of the cost of national defense is borne as a hidden or implicit tax by the draftee and the reluctant volunteer. The implicit tax on an individual draftee or reluctant volunteer is the maximum amount he would be willing to pay to buy his way out of military service. This amount, it may be noted, could be less than the difference between what he receives in service and the amount which would just induce him to volunteer for service in the absence of a draft.[2] The difference between civilian and service earnings also constitutes a real cost to society in the form of foregone output. In a price system, the aggregate implicit taxation implied by a draft system would be made into an extra wage bill to the Department of Defense, and an explicit tax on the general taxpayer.

Further analysis suggests further reduction of real costs in a volunteer system. A

* The author is assistant professor of economics at Brown University. He gratefully acknowledges helpful comments and suggestions from Gary S. Becker, Jacob Mincer, and a referee.

[1] For a discussion of the workings of the Selective Service System, see Fisher [5, pp. 72–3].

[2] For this distinction, the author wishes to thank an anonymous referee. For a detailed discussion, see Pauly and Willet, [9, pp. 53–7].

market solution would imply a different, presumably more efficient production process within the military. With the price of their less skilled labor input up, they would tend to employ less of it and substitute physical and human capital. Under the present system this labor input is, if not a free good to the military, at least a subsidized good, and this may be expected to reduce the incentive to economize in its use.

Another efficiency implication arises from the fact that men who enlist voluntarily are more likely to reenlist than draftees and draft-motivated enlistees. An increase in reenlistments would lower labor turnover costs, notably those incurred in the training of recruits. The better motivation of voluntary enlistees, as reflected in reenlistment patterns, might also be expected to have positive effects on productivity.

Less obvious, perhaps, than these explicitly specified effects are the effects of uncertainty associated with the present system. A changing world and changing deferment policies operate to keep many in a state of uncertainty as to whether or when they will be called. This distorts the pattern of investment in human capital decisions, and perhaps also the pattern of marriage and family planning decisions.[3] Of course, elimination of uncertainty does not require elimination of a draft system. A draft by lottery at an early age, say eighteen, as proposed by the Commission [12], would go far to eliminate uncertainty.

These, then are the real costs to society of the present system of military manpower procurement as opposed to a volunteer system: misallocation of resources, reduced productivity in civilian and military sectors, and distortion of the pattern of investment in human capital and family planning decisions.

In Section I we develop a model for the supply of enlisted volunteers for military service, and derive from it an equation that can be estimated statistically. In Section II we present and evaluate the results of least squares regression estimation, from time series and cross-section data, of the parameters of the equation. In Sections III and IV the results are used to determine the effect of the draft on enlistments, and the money cost to the Department of Defense and the general taxpayer of an all-volunteer force.

I. *A Model for the Supply of Enlisted Volunteers*

Enlistment in the Absence of a Draft

The Individual Decision to Enlist. In deciding whether or not to enlist in the absence of a draft, an individual is deciding between the returns to enlistment and the returns to some (best) civilian alternative. If he enlists, he may expect to receive a stream of earnings in the military W_{M1}, $W_{M2} \cdots W_{Mn}$ over periods $1, 2, \cdots n$ respectively. If he does not enlist, his (best) alternative is a stream of earnings in the civilian economy W_{C1}, W_{C2}, $\cdots$ W_{Cn}. The returns to enlistment would be the present value of the sum over periods 1 to n of earnings in M (the military), and the returns to nonenlistment would be the present value of the sum of earnings in C (the civilian sector). In deciding whether or not to enlist, then, the individual compares the present value of volunteering

$$V_M = \sum_{j=1}^{n} \frac{W_{M_j}}{(1+i)^j} \tag{1}$$

to the present value of not volunteering

$$V_C = \sum_{j=1}^{n} \frac{W_{C_j}}{(1+i)^j} \tag{2}$$

[3] A colleague, in Las Vegas on the day of the announcement that men married after midnight (that day) would not be deferred, observed the price of a girl for a quick marriage ceremony climb to two thousand dollars by late afternoon in the local bars.

Earnings in this discussion in principle include nonpecuniary, or psychic, as well as pecuniary components. The influence of the nonpecuniary on the supply of volunteers will be made explicit in the next section. At this point it is sufficient to think of it as included in the W_{M_j} and W_{C_j}.

Some modification of the present value stream V_M seems to be indicated by empirical considerations. Few servicemen reenlist, and fewer still remain in service for 20 years, the minimum necessary to qualify for a pension.[4] V_M then becomes approximately

$$\sum_{j=1}^{m} \frac{W_{M_j}}{(1+i)^j} + \sum_{j=m+1}^{n} \frac{W_{MC_j}}{(1+i)^j},$$

where W_{MC_j} is the jth period earnings in the civilian economy of an individual who served one term (periods 1, 2, $\cdots m$, $m<n$) in the military. If the time period units are years, V_M can be rewritten as

$$\sum_{j=1}^{3} \frac{W_{M_j}}{(1+i)^j} + \sum_{j=4}^{n} \frac{W_{MC_j}}{(1+i)^j}.$$

Some recent evidence suggests that $W_{MC_j} \simeq W_{C_j}$, i.e., that there is little difference in earnings of veterans and nonveterans.[5] If true, this leads to a simplification in the specification of the enlistment model. If

$$\sum_{j=4}^{n} \frac{W_{MC_j}}{(1+i)^j} \simeq \sum_{j=4}^{n} \frac{W_{C_j}}{(1+i)^j}$$

the enlistment decision would be based on a comparison of

$$\sum_{j=1}^{3} \frac{W_{M_j}}{(1+i)^j} \quad \text{with} \quad \sum_{j=1}^{3} \frac{W_{C_j}}{(1+i)^j}.$$

In comparing

$$\sum_{j} \frac{W_{M_j}}{(1+i)^j} \quad \text{and} \quad \sum_{j} \frac{W_{C_j}}{(1+i)^j},$$

note that discounting is important only if the time patterns of returns in M and C differ significantly. In fact, earnings increase monotonically over the first three years in both sectors, providing some justification for neglecting discounting over the three-year enlistment period. If discounting can be neglected over this relatively short period, it is possible to simplify still further. The decision to enlist would then be based on a comparison of

$$W_M = \sum_{j=1}^{3} W_{M_j} \quad \text{and} \quad W_C = \sum_{j=1}^{3} W_{C_j},$$

with enlistment taking place if $W_M > W_C$.

The Aggregate Enlistment Rate. The number of volunteers depends, therefore, on the distribution of civilian alternatives W_C. In order, then, to say something about the aggregate supply of volunteers, it is helpful to know something about the distribution of W_C.

Since W_C has been defined to include both money and psychic income, an individual can be said to enlist if $W_M > W_C + dW_C$, where W_M is redefined as money earnings in M,[6] W_C as money earnings in C, and d is a coefficient measuring the taste or relative preference for military service.

[4] Department of Defense. All data on military manpower and military pay are from the Department of Defense, unless otherwise noted.

[5] Gilman [6] reports the results of an attempt to compare earnings of veterans and nonveterans. Earnings are compared for veterans and nonveterans only in the same age and education classes. College graduates are not considered, but the omission may not be significant, as relatively few former enlisted men are in this class. In any case, Gilman's data, represented in age earnings profiles, shows a similar pattern for all education classes considered. Earnings of nonveterans are a few percentage points (perhaps as many as 12) higher than earnings of veterans at the youngest ages, but the differential steadily narrows until, at about age 45, it vanishes. Thereafter, earnings of veterans tend to be slightly above earnings of nonveterans.

[6] Money earnings in this formulation include all forms of material compensation such as the room and board provided most in the military, as distinguished from psychic compensation.

That is, if $d \neq 0$, an individual would not be indifferent between M and C if money earnings in M just equaled those in C, but would require some compensating differential. If, for example, he receives more psychic income in C, $d>0$, and he would enlist only if he receives money earnings in M greater at least by dW_C than those he could receive in his best alternative in C.

Now, what can be said about the distribution of civilian alternatives, i.e., the distribution of money earnings in $C(W_C)$ and the distribution of tastes for military service (d)? Some evidence suggests that the distribution of earnings is approximately lognormal, i.e., the distribution of earnings is unimodal and skewed to the right, and the distribution of the *log* of earnings is approximately normal.[7] Less is known about the distribution of relative preferences, or tastes for military jobs versus civilian. One source of information is the response of a large sample of male high school seniors in the Project Talent study to a question on attitudes toward a military career. The distribution of tastes for military service, as read from this response, is shown in Table 1.

TABLE 1—ATTITUDES TOWARD A MILITARY CAREER

Attitude	Percent
like very much	14.0
like fairly well	20.3
indifferent	23.1
dislike a little	12.0
dislike very much	30.6

Source: Project Talent: The American High School Student, University of Pittsburgh, 1964, p. K-16.[8]

[7] Aitchison and Brown cite studies indicating that "the more homogeneous the group of income recipients is the more likely is the lognormal curve to yield a good description of the income distribution; this is again more nearly true if the income is derived from a single source (if for example it consists entirely of earnings from employment)" [1, pp. 116–118].

[8] *Project Talent* [13] is a report on a study of the backgrounds, abilities, and educational and career plans of American high school students.

The percentage in the table may be interpreted as evidence on the distribution of d. Each attitude may be represented by a number (d) that, when multiplied by civilian earnings potential (W_C) gives the compensating differential (dW_C) that would induce enlistment. Thus, the attitude of "dislike very much" corresponds to a relatively high value of d, whereas at the other extreme "like very much" corresponds to a low, or even a negative, d-value.

It would seem, at a first reading, that the distribution of d-values among the respondents is bimodal, with one peak in the center and another in the upper tail. This interpretation may be modified by the following line of reasoning. The extreme upper tail option "dislike very much" is apparently associated with one mode of the distribution. This attitude, unlike the other ("indifferent") associated with a mode, may embrace widely varying d-values. For example, a payment by the military of perhaps twice their civilian wage might outweigh the unpleasantness of serving for some by enough to induce them to volunteer, whereas others could not be adequately compensated even with three or four times this amount. If there does exist some such distribution of d within the extreme class, this is an argument for representing the class by more than one point on the axis of tastes (d values). The distribution of d could then become unimodal and it would be skewed to the right.

Consider now the distribution of (W_C+dW_C), and observe that

(3) $W_C + dW_C = W_C(1 + d)$, and

(4) $ln\ [W_C(1 + d)] = ln\ W_C + ln(1 + d)$.[9]

As noted above, $ln\ W_C$ is approximately normally distributed. If d is unimodal and

[9] $Ln\ (1+d)$ is defined for $-1<d\leq\infty$. The restriction is however, not serious, as $d\leq -1$ would correspond to an individual volunteering for $W_M\leq 0$.

positively skewed, then $(1+d)$ is also, and $ln(1+d)$ would be more symmetrically distributed. We might assume it is approximately normally distributed. If W_C and d are not related, $ln[W_C(1+d)]=ln\ W_C+ln(1+d)$ would be normally distributed, since the sum of two independent normally distributed variables is itself normally distributed.

Any given value of W_M would cut off some part of the distribution of $W_C(1+d)$, or, any given value of $ln\ W_M$ would cut off some part of the distribution of $[ln\ W_C+ln(1+d)]$. It has been presumed earlier that all in the eligible population for whom $W_M>W_C\ (1+d)$ would enlist. More precisely now, the proportion of the eligible population that would enlist, which measures the aggregate supply curve to the military, is given by the area under the normal curve of $[ln\ W_C+ln(1+d)]$ to the left of $ln\ W_M$. Equivalently, the supply of enlistments is given by the value of the cumulative distribution function of $[ln\ W_C+ln\ (1+d)]$ at $ln\ W_M$.

Some inferences about the elasticity of this aggregate supply curve can be made. In a mathematical appendix, we show that the elasticity increases for $0<ln\ W_M<[ln\ W_C+ln(1+d)]^*$, where * denotes some point (that depends on μ and σ, the expected value and standard deviation of $[ln\ W_C+ln(1+d)]$) between $\mu/2$ and μ, and decreases for $ln\ W_M>[ln\ W_C+ln(1+d)]^*$.

For purposes of statistical estimation, we would like to write the enlistment rate as a simple function, consistent with the behavior of the elasticity function over a relevant range, of $ln\ W_M$, and the mean and standard deviation, μ and σ, of the distribution of $[ln\ W_C+ln(1+d)]$, the main determinants of turning points. In symbols we have equation (5),

$$\frac{E}{P}=f[ln\ W_M,\ \mu,\sigma] \tag{5}$$

where E=enlistment, P=population and E/P=the enlistment rate.

We first assume that $\sigma_{ln\ W_C+ln(1+d)}$, a measure of relative variation in the distribution of $W_C(1+d)$, has remained constant over time.

$$\begin{aligned}&\sigma^2_{ln\ W_C+ln\ (1+d)}\\ &\quad=\sigma^2_{ln\ W_C}+\sigma^2_{ln\ (1+d)}+2\sigma_{ln\ W_C\ ln\ (1+d)}\end{aligned} \tag{6}$$

Considering the terms in equation (6), note that the distribution of tastes $(\sigma^2_{ln(1+d)})$ is conventionally assumed constant for purposes of economic analysis, and the relative variation in the distribution of civilian earnings $(\sigma^2_{ln\ W_C})$ varied only slightly over the period of the empirical analysis. Again assuming W_C and d to be independent, the covariance term $(2\sigma_{ln\ W_C ln(1+d)})$ vanishes.

We do not assume that the mean of the distribution, $\mu_{ln\ W_C+ln(1+d)}$, has remained constant, but the expression may be simplified.

$$\mu_{ln\ W_C+ln\ (1+d)}=\mu_{ln\ W_C}+\mu_{ln\ (1+d)}. \tag{7}$$

Again assuming no change in the distribution of tastes, $\mu_{ln(1+d)}$, a measure of average preference for military service may be neglected leaving $\mu_{ln\ W_C}$ as a variable determinant of enlistment.

These simplifying assumptions make it possible to rewrite (5) as

$$\frac{E}{P}=f(ln\ W_M,\ \mu_{ln\ W_C}) \tag{8}$$

The mean of the lognormal distribution of $W_C(\mu_{ln\ W_C})$ is just equal to the median of the distribution. And since the operation of taking logarithms leaves the median of a distribution unchanged, the median of the distribution of $ln\ W_C$ is just equal to the logarithm of the median of the distribution of W_C. More concisely,

$$\mu_{ln\ W_C}=\widetilde{ln}W_C=ln\ \tilde{W}_C, \tag{9}$$

where $\sim$ denotes median, so (8) becomes

$$\frac{E}{P} = f(\ln W_M, \ln \tilde{W}_C). \tag{10}$$

We are now prepared to move from the general functional notation of (5), (8), and (10) above, to a specification of the functional form of the relation between E/P, $\ln W_M$, and $\ln \tilde{W}_C$. The specified function should be as simple as possible, consistent with the elasticity properties derived earlier. The linear function

$$\frac{E}{P} = \alpha' + \beta_1' \ln W_M + \beta_2' \ln \tilde{W}_C + \epsilon \tag{11}$$

in which ϵ contains random variations in tastes, etc., unrelated to $\ln W_M$ or $\ln \tilde{W}_C$, satisfies these conditions. Its simplicity needs no elaboration, and the resulting equation to be estimated will be shown further on to be consistent, over the relevant range, with the elasticity properties. However, in the form of equation (11), the linear function is consistent with the model from which it is derived only in the special case $\beta_2' = -\beta_1'$. To see why this is so, consider the model, in which the enlistment rate is determined by the level of military earnings vis-à-vis the distribution of civilian earnings. If W_M and $\tilde{W}_C$ are both increased by the same percentage amount, with the standard deviation unaffected, the enlistment rate is unaffected, since $\ln W_M$ and $\ln \tilde{W}_C$ both increase by the same absolute amount. Suppose W_M and $\tilde{W}_C$ are doubled. Then

$$\frac{E}{P} = \alpha' + \beta_1' \ln (2W_M) + \beta_2' \ln (2\tilde{W}_C) + \epsilon \tag{12}$$

$$= \alpha' + \beta_1' \ln W_M + \beta_2' \ln W_C + (\beta_1' + \beta_2') \ln 2 + \epsilon \tag{13}$$

E/P in (13) equals E/P in (11) only if $\beta_1' + \beta_2' = 0$, i.e., only if $\beta_2' = -\beta_1'$. Hence we rewrite (11) as

$$\frac{E}{P} = \alpha' + \beta_1' \ln W_M - \beta_1' \ln \tilde{W}_C + \epsilon \tag{14}$$

or

$$\frac{E}{P} = \alpha' + \beta_1' \ln \left(\frac{W_M}{\tilde{W}_C}\right) + \epsilon. \tag{15}$$

Some modification of (15) may be required to take account of an element of risk associated with earnings in the civilian economy: the chance of being unemployed. In principle $\tilde{W}_C$ could be median earnings of all workers, including the unemployed. If so, no modification would be required, but if $\tilde{W}_C$ is median earnings of full-time employed workers, then unemployment must be introduced explicitly. Expected Civilian earnings, adjusted for the chance of unemployment, could be represented as:

$$\tilde{W}_C^* = p_e\tilde{W}_C + p_uW_u, \tag{16}$$

a weighted average of earnings of full-time employed workers $\tilde{W}_C$ and earnings of unemployed workers W_u, the respective weights being p_e, the probability of being employed and p_u, the probability of being unemployed. Since both p_u and W_u are very small (W_u may be zero), the product p_uW_u makes a negligible contribution to $\tilde{W}_C^*$, and may be neglected in the expression for $\tilde{W}_C^*$. Equation (15) now becomes

$$\frac{E}{P} = \alpha' + \beta_1' \ln \left(\frac{W_M}{\tilde{W}_C^*}\right) + \epsilon \tag{17}$$

where $\tilde{W}_C^* = p_e\tilde{W}_C$.

Enlistment with a Draft

In the preceding sections we have specified a model of the individual decision to enlist and the aggregate enlistment rate in a free labor market, i.e., in a market with no compulsory military service or draft system. In this section the draft is introduced as an additional determinant of the supply of volunteers for military service.

It is introduced explicitly through its effect on expected earnings in the civilian economy, or more correctly, expected earnings associated with a decision not to enlist. It will be recalled that $\tilde{W}_C^* = p_e \tilde{W}_C$ has been defined as expected earnings in the civilian economy in the absence of a draft. But if some proportion of those who elect to remain in the civilian economy will be taken into the military, then expected earnings in the civilian economy are a weighted average of civilian and military earnings. The expected wage associated with a decision not to enlist may be represented as:

$$\tilde{W}_C' = p_c \tilde{W}_C^* + p_d W_M, \tag{18}$$

a weighted average of expected civilian earnings $\tilde{W}_C^*$ and military earnings W_M, the respective weights being p_c, the probability of remaining a civilian, and p_d, the probability of being taken into the military.[10]

Equation (17) of the last section can be rewritten taking account of the draft, using (18) above, as:

$$\frac{E}{P} = \alpha' + \beta_1' \, ln \left(\frac{W_M}{\tilde{W}_C'} \right) + \epsilon \tag{19}$$

$$= \alpha' + \beta_1' \, ln \left(\frac{W_M}{p_c \tilde{W}_C^* + p_d W_M} \right) + \epsilon \tag{20}$$

$$= \alpha' + \beta_1' \, ln \left(\frac{W_M}{p_c p_e \tilde{W}_C + p_d W_M} \right) + \epsilon \tag{21}$$

Some manipulation of (21) yields

$$\frac{E}{P} = \alpha' - \beta_1' \cdot \left(ln \, p_c + ln \, p_e + ln \frac{\tilde{W}_C}{W_M} \right) + \epsilon.^{11} \tag{22}$$

[10] The implicit assumption is made that the draftee and the volunteer receive the same compensation (W_M).

[11] Full details are given in Fisher [5, p. 24].

To make this expression operational, p_c and p_e must be defined in terms of observable quantities. Let

$$p_c = \delta_1 \left(1 - \frac{A}{P} \right)^{\delta_2}. \tag{23}$$

where A = enlistments plus inductions, or total accessions into the military (in a given period) and P = the eligible population, as in the denominator of the dependent variable. A/P, the "accession rate," is a measure of military demand in the civilian labor market. p_c, the chance of remaining a civilian, is supposed to be a function of this demand. This, it may be noted, is a more general formulation than would be obtained by setting $p_c = 1 - A/P$, i.e., by setting $\delta_1 = \delta_2 = 1$. Thus, p_c may represent some longer run concept than the current level of military demand, but as a first approximation, at least, is some function of this demand. One might object that inductions, and not total accessions, should be used to measure the effect of the draft on potential enlistees. But use of the induction rate could lead to bias in the estimation procedure. Suppose conditions of military service become more attractive. The enlistment rate will rise, but at a given level of manpower strength the induction rate must fall. Thus, enlistments could be spuriously negatively related to inductions. The same result follows from a worsening of conditions of service; the enlistment rate falls, but at a given level of strength inductions rise. Any coefficient obtained to measure the relationship between inductions and enlistments is likely then to be an underestimate of that relationship. Inductions may indeed affect enlistments, but enlistments may, as indicated, affect inductions. The desired level of manpower, the accession rate, is however determined exogenously by the Department of Defense and enlistments (and inductions) may then be affected.

Similarly, let

$$p_e = \gamma_1(1 - U)^{\gamma_2}, \tag{24}$$

where U=the unemployment rate in the eligible population.

Substituting (23) and (24) in (22) yields, upon further manipulation,

$$\frac{E}{P} = \alpha + \beta_1 \ln \frac{\bar{W}_C}{W_M} + \beta_2 \ln (1 - U) + \beta_3 \ln\left(1 - \frac{A}{P}\right) + \epsilon.^{12} \tag{25}$$

The possibility of least squares bias in the estimation of the parameters of (25), treated above for accessions, also arises in connection with earnings and unemployment. To the extent it exists, it may be removed by the use of lagged values to measure earnings and unemployment. The use of lags is suggested, further on, by other considerations as well.

To obtain elasticities from the coefficients in equation (25), a simple transformation is needed. By definition,

$$\text{elasticity } \xi = \frac{d(\ln y)}{d(\ln x)} \tag{26}$$

$$= \frac{\frac{dy}{y}}{\frac{dx}{x}}. \tag{27}$$

If $y=a+b \ln x$, the coefficient (b) equals $dy/d(\ln x)$. Hence, the elasticity is given by

$$b \cdot \frac{1}{y} = \frac{dy}{d(\ln x)} \cdot \frac{1}{y} \tag{28}$$

$$= \frac{d(\ln y)}{d(\ln x)} \tag{29}$$

Note that the elasticity is not constant, but varies inversely with y; the smaller y, the larger the elasticity, the larger y, the smaller the elasticity.

Recalling our earlier discussion of the behavior of the elasticity function at various points along a supply curve derived from normal distributions of civilian earnings and tastes for military service, it is seen that the (semi-log) linear approximation (eq. 11) is valid if $\ln W_M$ is greater than $\ln \bar{W}_C$, and also if $\ln W_M$ is not very much less than $\ln \bar{W}_C$; i.e., the approximation is valid if $\ln W_M > \ln \bar{W}_C/2 + \Delta$, where Δ depends on the mean and standard deviation of $\ln W_C$.[13] The inequality is satisfied over the range of projected observations.[14]

II. *Empirical Results*

Time Series: The Data

Time series data are used to estimate the parameters of equation (25), derived at the end of the preceding section. In a simple least squares regression, the following empirical measures of the specified variables are used:

(a) $\frac{E}{P}$, the Enlistment Rate

Potential enlistees are given an aptitude test, and on the basis of this test distributed into five "mental categories." No enlistments are accepted from category V, the lowest, and only a few, determined by demand, from category IV, the next lowest. There is no official limitation on enlistments from categories I, II and III, which are then used in our estimation procedure. The population base for enlistments is the population of male civilians, age 17–20, from which 85–90 percent of

[12] Full details are given in Fisher [5, p. 26].

[13] This statement is strictly correct only for $\mu_{\ln(1+d)}=0$, or alternatively, for $\mu_d=0$. Evidence on the distribution of d from Table 1 above suggests that these equalities are at least approximately satisfied.

[14] At the beginning of the time series (3rd quarter 1957), W_C was $2,333 (annually) and W_M $2,290. At the end of the series (3rd quarter 1965), the figures were, respectively, $3,292 and $2,628. With the end of the draft, projected values for W_M and $\bar{W}_C$ involve an increase in W_M relative to $\bar{W}_C$ such that $W_M > \bar{W}_C$. Sources for these figures and projections are discussed in Sections II and IV below.

enlistments are drawn. The population series is from *Current Population Reports*, (*CPR*), which present annual observations. Estimates of quarterly observations are obtained by linear interpolation.

(b) W_M, Military Earnings

Earnings of first-termers in the military have varied very little over the period of analysis, and indeed, over a period stretching back to World War II. Two or three small increases, in recent years (1964, 1965), have imparted an upward trend to an otherwise unvarying series. W_M includes basic pay, quarters and subsistence allowances, and an imputed value of medical services.

(c) $\tilde{W}_C$, Median Civilian Earnings

Median civilian earnings are represented by the median income of year round, full time male workers, ages 14–19 and 20–24. While median earnings would be preferable, the lack of an appropriate earnings series is probably not significant, as the income of the above includes relatively little property income. Of course, the youngest enlistees are 17, so the earnings of 14–16 year-olds are not immediately relevant to an enlistment decision. But since income of only year round, full-time workers is reported, and most 14–16 year-olds are in school, the incomes of 17–19 year-olds probably dominate the incomes of year-round, full time workers age 14–19. For a typical enlistment, say into the Army at age 18, an individual would be foregoing about two years of civilian earnings before age 20 and a year of civilian earnings after age 20. Accordingly a weight of $\frac{2}{3}$ is given to 14–19 and $\frac{1}{3}$ to 20–24, in constructing each observation on $\tilde{W}_C$. The income series are from *CPR*, which present annual observations. Estimates of quarterly observations are obtained by linear interpolation.

(d) U, the Unemployment Rate

The unemployment rate is the rate for all males, age 18–19. This series is from unpublished Bureau of Labor Statistics figures.

(e) $\frac{A}{P'}$, the Accession Rate

Since the accession rate is used to measure the effect of the draft on enlistments, accessions are related to the population of potential enlistees—the population used with enlistments to give enlistment rates.

The period of analysis runs from the third quarter of 1957 through the fourth quarter of 1965. The period before fiscal 1958 (third quarter calendar 1957) was passed over due to a quota system in the few years preceding fiscal 1958 which required minimum percentages of category IV's accepted by the services. If enforced, and evidence from enlistments rates indicates that it was, this system might well have resulted in rejection of volunteers from categories I, II, and III, or at least in their less active recruitment. Another reason for excluding the period before 1957 is that the Navy drafted about 30,000 men in fiscal 1956, obscuring the enlistment picture for that year.

Quarterly, rather than monthly or yearly observations, are taken. Since the period considered extends over only eight and one-half years, and the model specifies three independent variables, clearly annual observations would provide too few degrees of freedom for reliable estimation. Monthly observations, on the other hand, might be subject to excessive variation from random and institutional sources, introducing erratic factors into the model. Dummy variables are introduced to adjust the model to seasonal patterns in the flow of enlistments. A dummy variable for each season is introduced, the most important of these being the summer dummy due to the heavy influx of new high school graduates over the summer months.

Time Series: Results

Results of the estimation of the parameters of equation (25) are presented, with some explanation, in equation (30) below.

$$
\begin{aligned}
\frac{E}{P} = {} & .00751 - \underset{(.00324)}{.00709}\, ln\left(\frac{\bar{W}_C}{W_M}\right) \\
& - \underset{(.01018)}{.00891}\, ln\,(1 - U) \\
& - \underset{(.041)}{.312}\, ln\left(1 - \frac{A}{P}\right) \qquad (30) \\
& - \underset{(.00069)}{.00133}\, SP + \underset{(.00065)}{.00254}\, SU \\
& - \underset{(.00056)}{.00196}\, A \qquad R^2 = .90
\end{aligned}
$$

(a) $ln(\bar{W}_C/W_M)$ and $ln(1-U)$ are lagged one period.
(b) SP, SU, and A are seasonal dummy variables for spring, summer and autumn respectively.
(c) The figures in parentheses are standard errors.

Using equation (28), we may calculate elasticities from equation (30). For example, the elasticity of enlistment response (at the mean enlistment rate) to earnings is about $-.46$.

The use of lagged earnings and employment variables is suggested by both theoretical and data considerations. Since some time may be required to adjust to changing conditions affecting the enlistment decision and the period units of analysis (quarters) are fairly short, it may be, for example, that enlistments in period (t) depend on earnings in period ($t-1$). Earnings, however, tend to increase monotonically and slightly from period to period, making earnings in period (t) nearly the same as earnings in period ($t-1$). Of course, this behavior of the earnings series is partly due to our treatment of the data, in which quarterly observations are obtained by interpolation from an annual series. The annual observations themselves tend to increase moderately though not, of course, as regularly as the interpolated observations. Hence, we are not surprised to find virtually no difference in size or significance of the regression coefficients associated with lagged and current earnings. Unemployment rates are more volatile, and unemployment in the lagged form is found to bear a closer relation to enlistments. Again, this may be explained by the time taken to adjust to changing labor market conditions, as well as the possible least squares bias noted earlier.

Before calculating from equation (30), the effect of the draft on enlistments and the money cost to the Department of Defense of ending the draft, we develop from cross section data some corroborating estimates of supply parameters.

Cross-Section

In [2], Altman estimates, from cross-section data, parameters of an equation similar in some respects to (25).

$$
\begin{aligned}
ln\, R = {} & .493 + \underset{(3.66)}{.379}\, ln\, Y \\
& + \underset{(4.02)}{.190}\, ln\, U - \underset{(-3.85)}{.101}\, ln\, N \qquad (31) \\
& \qquad R^2 = .86
\end{aligned}
$$

in which R = the enlistment rate, Y = the ratio of military earnings to (average) civilian earnings, U = the unemployment rate, and N = the percentage of nonwhite males. The figures in parentheses are t-values.

In an attempt to take account of draft pressure, Altman uses an attitude survey to separate out the mainly draft-induced from all (I–III) enlistees. Another equation, relating the same independent variables to "genuine" enlistments only, is estimated as

$$\ln Z = -.807 + \underset{(2.13)}{.813}\, Y + \underset{(1.97)}{.344}\, U - \underset{(-.04)}{.004}\, N \quad R^2 = .69 \tag{32}$$

Since the coefficients of (31) and (32) are (constant) elasticities, it is seen that the earnings elasticity in (31) is close to that in (30), and both are below that in (32). The elasticities in (30) and (31) are for earnings with respect to enlistments *with* a draft; in (32), for earnings with respect to enlistments *without* a draft. As noted earlier, we intend to use equation (30) to derive an estimate, based on observed behavior, of the effect of the draft on enlistments, and in the same process, of the level of military earnings required to attract any number of volunteers. But first, using the attitude survey result of 38 percent draft-induced enlistees, we calculate from equations (28) and (30) an earnings elasticity at the mean enlistment rate without a draft, as about $-.74$. This is close to the $-.81$ in (32), so that time series and cross-section estimates of the effect of changes in earnings on enlistments, with and without a draft, are in agreement.

A property of equation (25) noted earlier is that elasticity would be relatively large with a relatively small proportion of the population already enlisted, and relatively small with a relatively large proportion of the population already enlisted. Thus, elasticity of enlistments without a draft is (absolutely) larger $(-.74)$ than elasticity of enlistments with a draft $(-.46)$, since the proportion of the population enlisted is smaller.

An alternative form for a supply equation with this property is developed in [2] and [7]. This "complement" equation, in which elasticity $\xi = -b(1-Z/Z)$, where b is the regression coefficient and Z is the enlistment rate, is estimated from cross-section data as

$$\ln(1-Z) = 5.03 - \underset{(-3.59)}{.109}\ln Y - \underset{(-1.90)}{.035}\ln U \quad R^2 = .73 \tag{33}$$

Earnings elasticities with and without a draft are .44 and .81 respectively, in agreement with both cross-section constant elasticities and time series elasticities.

A further test for the model and equation of Section I is provided by the cross-section data. In addition to the above comparison of time series and cross-section results, we apply equation (25) to the cross-section data, and estimate

$$\frac{E}{P} = -.105 - \underset{(-3.28)}{.098}\ln\left(\frac{W_C}{W_M}\right) - \underset{(-1.47)}{.226}\ln(1-U) \qquad R^2 = .67 \tag{34}$$

The significance of the coefficients, as measured by the t-values in parentheses, and the magnitude of the coefficient of determination, are nearly the same as for the complement equation (33). Earnings elasticities with and without a draft are $-.49$ and $-.82$ respectively, in agreement with all other estimates.

III. *The Effect of the Draft on Enlistments*

Evidence from an attitude survey has been cited as suggesting that as of 1964 about 38 percent of all volunteers are draft-induced. We now use equation (30), estimated from time series data, to calculate the percentage of draft-induced volunteers. We do this by using the equation to predict the effect on the enlistment rate of a change in the induction rate. Recall

$$\frac{E}{P} = a + b_1 \ln\left(\frac{\bar{W}_C}{W_M}\right) + b_2 \ln(1-U) + b_3 \ln\left(1 - \frac{A}{P}\right), \tag{35}$$

where $b_1<0$, $b_2<0$ and $b_3<0$.
Then

$$\frac{E}{P} = a + b_1 \ln\left(\frac{\tilde{W}_C}{W_M}\right) + b_2 \ln(1-U) - b_3\left(\frac{A}{P}\right) \tag{36}$$

since

$$\ln\left(1-\frac{A}{P}\right) \simeq -\frac{A}{P}$$

$$\text{for } \frac{A}{P} \text{ close to } 0.$$

$$\frac{E}{P} = a + b_1 \ln\left(\frac{\tilde{W}_C}{W_M}\right) + b_2 \ln(1-U) - b_3\left(\frac{E}{P}+\frac{I}{P}\right)^{15} \tag{37}$$

$$\frac{E}{P} = \frac{a}{1+b_3} + \frac{b_1}{1+b_3}\ln\left(\frac{\tilde{W}_C}{W_M}\right) + \frac{b_2}{1+b_3}\ln(1-U) - \frac{b_3}{1+b_3}\left(\frac{I}{P}\right). \tag{38}$$

Suppose the induction rate I/P is set equal to zero, its value if inductions were eliminated. The enlistment rate E/P in equation (38) is reduced by the amount of the term $[(-b_3/1+b_3)(I/P)]$, which drops out of the equation. The average enlistment rate is 0.153 and the enlistment rate calculated from equation (38) (with $[(-b_3/1+b_3)(I/P)]=0$) is .0117, a reduction of about 24 per cent.

This is less than perfectly consistent with the 38 percent reduction implied by the attitude survey. Possible explanations of the apparent discrepancy are: (1) Attitude survey response is in general unreliable and may here lead to an overestimate of the effect of the draft on enlistments. Error typically accompanies the reporting even of items, such as consumer income, capable of exact evaluation and representation. It might be expected, then, that response to a subtle question of motivation, and for an action (the decision to enlist) taken perhaps several years prior to the time of the question, is likely to be still more subject to error, still less reliable. It is conceivable, for example, that the draft (as motivation for enlistment) represents a convenient catch-all or excuse. This would be especially true for those unable to realize original hopes or objectives in enlisting, a failure perhaps difficult to admit. (2) Prediction from a regression line is associated with increasing error as the difference between the projected value of the independent variable and its mean increases.[16] Thus, the coefficient $-b_3/1+b_3$ may reliably predict the effect on the enlistment rate of small movements around the mean induction rate, but not of a change of 100 percent (drop to zero) in the rate. Nothing is implied about the direction of error, but caution is suggested in interpreting the predicted enlistment rate. (3) While a drop to zero in the induction rate could be and has been interpreted as elimination of the service obligation, legal elimination might have a stronger effect on the enlistment rate. An induction rate equal to zero could leave some potential enlistees uninfluenced (at least for some time) by the draft, but others might prefer to enlist, given the persistence of the obligation.[17]

[15] We ignore the relatively (in comparison with I–III enlistments (E/P) and inductions (I/P)) small number of mental group IV enlistments.

[16] $S_p=S_u\sqrt{1+1/n+(x_p-\bar{x})^2/\sum(x-\bar{x})^2}$, where S_p is the standard error of prediction, S_u is the standard deviation in the distribution of the residual, x_p is the projected value of the independent variable, and $\bar{x}$, its mean.

[17] This distinction would become less important if the induction rate continued over many periods to remain at zero.

IV. *The Cost of Eliminating the Draft*

As indicated, an induction rate (in equation (38)) equal to zero would imply some decrease (24 percent) in the enlistment rate. The question now emerges: what would be required, in the absence of a draft, to generate the desired enlistment rate? More specifically, what increase in military earnings would be required to generate a voluntary enlistment rate sufficient to maintain desired levels of Armed Forces strength? An answer to this question is obtained from equation (38) by substituting the desired value of E/P, along with projected values of $\tilde{W}_C$, U, and I/P, and solving for W_M.

As a first approximation to the desired value of E/P, the accession rate A/P, which measures total inflow of genuine volunteers, draft-induced volunteers and draftees, might be suggested. But one source of reduced costs in a volunteer system would be improved retention. Higher reenlistment rates imply not only a reduction in resources currently employed in training activities, as noted in the induction, but also a lower rate of accessions to maintain force levels. Oi [7] cites Air Force studies showing higher reenlistment rates for genuine (as determined by attitude survey response) than for draft-induced volunteers. Even if attitude survey response is considered unreliable, it is nonetheless true that reenlistment rates are higher for all (genuine and draft-induced) volunteers than for draftees. On the basis of these differences in reenlistment rates Oi has calculated that, to maintain a force of 2.65 million men (approximately the level for 1957–65 and presumably the desired peace-time level), accessions in an all-volunteer system need only be about seven-tenths (.707) of accessions in a draft system. The enlistment rate substituted into equation (38) is then the average accession rate for 1957–65 adjusted by a factor of .707 to reflect the expected change in reenlistment rates in a volunteer system.

Median civilian earnings have been rising over time, so the latest value in the series is more relevant than, say, the average in determining the effect of any action, such as elimination of the draft, taken now or in the future. Accordingly the latest observation on $\tilde{W}_C$ (end of 1965) is substituted into equation (38). The unemployment rate, on the other hand, fluctuated considerably over the period 1957–65, and in the absence of a clear trend, the average rate is the projected rate. The induction rate is set equal to zero, as enlistments are no longer motivated by inductions.

Substituting the indicated values for E/P, I/P, $\tilde{W}_C$, and U into equation (38), we calculate the required increase in W_M as \$2105 per year (above W_M in late 1965). In a military force of 2.65 million men, about 2.3 million are enlisted, and of these, about half, or some 1.15 million, are first-termers. If the increase were to go to first-termers alone, the extra wage bill would come to \$2.42075 billion (annually). Yet, pay raises of this magnitude (\$2105 per year) would result in a wage structure in which first-termers were earning more than many trained experienced servicemen. To forestall probable associated diseconomies (low morale, low reenlistment rates) the increase could go to all enlisted men, thereby preserving the structure of differentials. The extra wage bill would then simply be twice the increment to first-termers, or \$4.8415 billion.

The absolute value of the earnings-elasticity at the enlistment rate associated with current military earnings in equation (38) is about .60. At the rate associated with military earnings required to maintain a force of 2.65 million men, it is about .43. The rate of decrease would appear to be moderate, suggesting that marginal increases in desired size of force would not

require increases in military earnings of an order of magnitude much greater than indicated above.

Modifications of the calculated wage increase are, of course, possible, given changes in the values of either or both of the independent variables ($\tilde{W}_C$ and U). For example, while the unemployment rate for 18–19 year-old males averaged around 16 percent over the period 1957–1965, it dropped to near 10 per cent in the fourth quarter of 1965. Accordingly, an unemployment rate of 10 percent can be substituted into the equation, making the required increase in military earnings \$2489 per year and the extra wage bill \$5.7247 billion. Or, suppose civilian earnings continue to increase. To measure the sensitivity of the required change in military earnings to an increase in civilian earnings, let W_C rise from \$10,000 (over the three-year enlistment period) to \$11,000, probably about right as of the fourth quarter of 1967. Then, the required increase in W_M is \$2579 per year, and the extra wage bill \$5.9317 billion. If both unemployment and (civilian) earnings change in the directions and magnitudes indicated, the required increase in military earnings is \$3000 per year and the extra wage bill \$6.9 billion.

The effect of a given percentage change in civilian earnings is considerably greater than the effect of the same percentage change in unemployment. This is a consequence of the relatively weak relation between unemployment and enlistments. If the estimated relation is for some reason an underestimate, as suggested perhaps by the cross-section findings, the required increase in military earnings (and the extra wage bill) associated with a fall in unemployment would be greater than indicated above.

To sum up, given the projected values for enlistments, population, earnings and unemployment, the additional money cost to the Department of Defense of an all-volunteer force would be not more than five to seven and a half billion dollars annually.[18] Recalling the introductory arguments it may be seen why five to seven and a half billion dollars is considered an upper bound. A more economical use of resources by the military, and a reduction in their training costs, both implied in a volunteer system, would reduce the projected tax increase. Increased output and earnings in the civilian sector, also implied in a volunteer system, would benefit some young men directly (the increased earnings), and the general taxpayer indirectly, through tax revenues generated by the increased earnings.

Appendix

Let

$$x = ln\, W_M$$

and the cumulative distribution function = the enlistment rate

$$= y = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{x} e^{-\frac{1}{2}((t-\mu)/\sigma)^2} dt.$$

Define

$$\xi = \frac{dy}{dx} \cdot \frac{x}{y}$$

where ξ = elasticity of y with respect to x. We are interested in the behavior of ξ at various points along the curve, i.e., in the effect on ξ of changes in x. By differentiation

$$(1) \quad \frac{\partial \xi}{\partial x} = \frac{y \cdot \dfrac{dy}{dx} + x \cdot \dfrac{d^2 y}{dx^2} \cdot y - x \cdot \left(\dfrac{dy}{dx}\right)^2}{y^2}$$

Since $y^2 > 0$, only the numerator of (1) need be considered in determining whether ξ decreases, is unaffected, or increases with increases in x.

[18] Fechter [3, p. 30] estimates the additional cost of attracting officers in the absence of a draft as \$.42 billion.

$$(4)\quad \text{Num}\,\frac{\partial\xi}{\partial x}=\frac{1}{2\pi\sigma^2}\left[(e^{-\frac{1}{2}z^2}-xze^{-\frac{1}{2}z^2})\int_{-\infty}^{x}e^{-\frac{1}{2}((t-\mu)/\sigma)^2}dt-xe^{-z^2}\right]\quad\text{where}\quad z=\left(\frac{x-\mu}{\sigma}\right)$$

and

$$(5)\quad \text{Num}\,\frac{\partial\xi}{\partial x}=\frac{1}{2\pi\sigma^2}e^{-\frac{1}{2}z^2}\left[(1-xz)\int_{-\infty}^{x}e^{-\frac{1}{2}((t-\mu)/\sigma)^2}dt-xe^{-\frac{1}{2}z^2}\right].$$

$$(6)\qquad G(x)=\left(1-\frac{x(x-\mu)}{\sigma}\right)\int_{-\infty}^{x}e^{-\frac{1}{2}((t-\mu)/\sigma)^2}dt-xe^{-\frac{1}{2}((x-\mu)/\sigma)^2}.$$

$$(2)\quad \frac{dy}{dx}=\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2}((x-\mu)/\sigma)^2}$$

$$(3)\quad \frac{d^2y}{dx^2}=\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2}((x-\mu)/\sigma)^2}\left(-\left(\frac{x-\mu}{\sigma}\right)\right)$$

Substituting (2) and (3) in (1), we get (4) and (5).

$$\text{Since}\qquad \frac{1}{2\pi\sigma^2}e^{-\frac{1}{2}z^2}>0,\qquad \frac{\partial\xi}{\partial x}\gtreqless 0,$$

$$\text{as}\qquad G(x)\gtreqless 0.$$

We show that $G(x)$, is positive at $x=0$, increases to a maximum at $x=\frac{\mu}{2}$, and becomes negative at $\frac{\mu}{2}<x<\mu$. If $x=0$,

$$(7)\qquad G(0)=\int_{-\infty}^{0}e^{-\frac{1}{2}((t-\mu)/\sigma)^2}dt>0.$$

If $x=\mu$,

$$(8)\qquad G(\mu)=\int_{-\infty}^{\mu}e^{-\frac{1}{2}((t-\mu)/\sigma)^2}dt-\mu$$

Let $s=\left(\dfrac{t-\mu}{\sigma}\right)$. Then

$$(9)\qquad G(\mu)=\int_{-\infty}^{0}e^{-\frac{1}{2}s^2}\sigma ds-\mu$$

$$(10)\qquad =\sqrt{\frac{\pi}{2}}\,\sigma-\mu\gtreqless 0\qquad \text{as}\ \frac{\mu}{\sigma}\lesseqgtr\sqrt{\frac{\pi}{2}}$$

Since $\mu/\sigma>\sqrt{\pi/2}$ in the distribution of civilian earnings of relatively homogeneous earners, young men of military age,[19] $G(x)$ is negative at $x=\mu$.

We have evaluated the function $G(x)$ for various values of x. We now consider more general behavior of $G(x)$.

$$(11)\qquad \begin{aligned}\frac{\partial G}{\partial x}&=\left(\frac{-2x+\mu}{\sigma}\right)\int_{-\infty}^{x}e^{-\frac{1}{2}((t-\mu)/\sigma)^2}dt\\&\quad+\left(1-\frac{x(x-\mu)}{\sigma}\right)e^{-\frac{1}{2}((x-\mu)/\sigma)^2}\\&\quad-e^{-\frac{1}{2}((x-\mu)/\sigma)^2}\\&\quad+x\left(\frac{x-\mu}{\sigma}\right)e^{-\frac{1}{2}((x-\mu)/\sigma)^2}\end{aligned}$$

$$(12)\qquad \begin{aligned}&=\frac{1}{\sigma}(\mu-2x)\int_{-\infty}^{x}e^{-\frac{1}{2}((t-\mu)/\sigma)^2}dt\\&\gtreqless 0\qquad \text{as}\ x\lesseqgtr\frac{\mu}{2}.\end{aligned}$$

Since $\partial\xi/\partial x$ behaves like $G(x)$, in particular num $\partial\xi/\partial x=g(x)\cdot G(x)$, where $g(x)=(1/2\pi\sigma^2)e^{-\frac{1}{2}z^2}$, ξ increases for $0<x<\mu/2$, continues to increase for $x>\mu/2$ until some x in the interval from $\mu/2$ to μ, and decreases as x increases beyond this point.

[19] From figures compiled at the Institute for Defense Analyses by J. House, $\mu\simeq\$3500$ and the difference between 80th and 20th percentile earnings$\simeq\$1800$. Assuming a distribution approximately normal over this range, $1.68\ \sigma=\$1800$, so $\sigma=\$1071$ and $\mu/\sigma=\$3500/\$1071>\sqrt{\pi}/2$.

REFERENCES

1. J. AITCHISON AND J. A. C. BROWN, *The Lognormal Distribution*. Cambridge 1957.
2. S. H. ALTMAN, "Earnings, Unemployment and the Supply of Enlisted Volunteers," unpublished manuscript.
3. S. H. ALTMAN AND A. E. FECHTER, "The Supply of Military Personnel in the Absence of a Draft," *Am. Econ. Rev.*, Proc., May 1967, *57*, 19–31.
4. G. S. BECKER, *Human Capital: A Theoretical and Empirical Analyses, With Special Reference to Education*. New York 1964.
5. A. C. FISHER, *The Supply of Enlisted Volunteers for Military Service*, unpublished Ph.D. dissertation, Columbia, 1968.
6. H. J. GILMAN, unpublished manuscript.
7. W. Y. OI, "The Economic Cost of the Draft," *Am. Econ. Rev.*, Proc., May 1967, *57*, 39–62.
8. ———, "The Costs and Implications of an All-Volunteer Force," in S. Tax, ed., *The Draft*, Chicago 1967.
9. M. V. PAULY AND T. S. WILLET, "Who Bears the Burden of National Defense?" in J. C. Miller, III, ed., *Why the Draft*, Baltimore 1968.
10. B. A. WEISBROD AND W. L. HANSEN, "Economics of the Military Draft," *Quart. Jour. Econ.*, August 1967, *81*, 395–421.
11. *Civilian Advisory Panel on Military Manpower Procurement, Report to the Committee on Armed Services of the House of Representatives*, Washington 1967.
12. *In Pursuit of Equity: Who Serves When Not All Serve? Report of the National Advisory Commission on Selective Service*, Washington 1967.
13. *Project Talent: The American High School Student*, University of Pittsburgh 1964.

Behavior of the Firm Under Regulatory Constraint

By Akira Takayama*

A recent article in this *Review* by Averch and Johnson [3] on the fair rate of return regulation in public utilities industries (or more generally "regulated industries") attracted the attention of many economists. In essence, Averch-Johnson argues that the introduction of an "active" constraint of a fair rate of return type would induce the firm to invest more than the original profit maximizing value of capital, and this would create an inefficient allocation of inputs. In addition to frequent citations in verbal arguments, it has invited some theoretical extensions by economists such as Westfield [8] and Klevorick [5]. All of the writers seem to accept the basic theoretical foundations set by Averch and Johnson. For example, Klevorick [5] recently proposed that the fair rate of return should be inversely related to the size of the capital of a firm. The intuitive basis for this is apparent, for it in effect creates a penalty as the amount of capital used by the firm increases.[1]

However, it seems to me that Averch-Johnson's original arguments are sketchy and contain somewhat ambiguous logic at several important points. Although their conclusion, for which they tried to find empirical support, that a firm will tend to increase its investment with the introduction of an active constraint seems to be correct, it does not follow from their arguments. In particular, there seems to be a serious error which arises from confusing movements along a curve and shifts of a curve. In Section II, we clarify those ambiguities, and if necessary, correct the mistakes. In the third section of this paper, we shall provide a new formulation of the problem. We shall find that the new formulation will give the same answer as Averch-Johnson, namely that a firm will tend to invest more with the introduction of an active constraint. It should be noted however that while the conclusion happens to be the same, this was not the necessity, for we needed a completely new formulation to obtain the old conclusion.

I *The Model*

Consider a monopoly producing a homogeneous output (Y) using two inputs, capital (K) and labor (L). The production function of the firm can be written:[2]

$$(1)\quad Y = F(L, K) \quad \text{where} \quad F_L > 0, \quad F_K > 0$$

The inverse demand function can be written:

$$(2)\quad p = p(Y) \quad \text{where} \quad p'(Y) < 0$$

* The paper was originally written in the summer of 1966. The author is professor of economics at Purdue University. He is grateful to Professors D. Emery, K. R. Kadiyala and H. Sonnenschein and to the referee for their comments. All possible mistakes are, of course, his.

[1] Westfield [8] argues that "... under a number of plausible circumstances higher price tags on capital goods leave the buyer either with undiminished or even with higher profits," pp. 440–1. This conclusion also seems to be an intuitively natural consequence of the Averch-Johnson result.

[2] Throughout this paper, the following notation will be used:

$$\frac{\partial}{\partial L} F(L, K) = F_L, \qquad \frac{\partial}{\partial K} F(L, K) = F_K,$$

$$\frac{\partial^2}{\partial L \partial K} F(L, K) = F_{LK},$$

$$\frac{\partial^2}{\partial K^2} F(L, K) = F_{KK}, \qquad \frac{\partial}{\partial L} G(L, K) = G_L,$$

$$\frac{\partial^2}{\partial L^2} G(L, K) = G_{LL}, \quad \text{etc.}$$

Assume that the firm has no control over the price of inputs (w for labor and r for capital). Profit (π), which is a function of L and K, can now be written as:

$$\pi = pY - wL - rK \tag{3}$$

Averch-Johnson and Klevorick require that pY be a concave function with respect to L and K.[3] Letting s be the "fair rate of return," the regulatory constraint can be written as:[4]

$$\frac{pY - wL}{K} \leqq s \tag{4}$$

Or:

$$g(L, K) \equiv -(pY - wL - sK) \geqq 0 \tag{4'}$$

If $s<r$, the firm would withdraw from the market. To sharpen the analysis we shall avoid the case $s=r$, and assume $s>r$. The problem is to maximize the profit function,[5] (3), subject to (4) and $L\geqq 0$, and $K>0$. The Lagrangian can be defined as:

$$\Phi(L, K, \lambda) = pY - wL - rK - \lambda(pY - wL - sK) \tag{5}$$

If the constraint function $g(L, K)$ is a concave function, then due to the concavity of the profit function, the following Kuhn-Tucker-Lagrange conditions are necessary and sufficient for a maximum.[6] However, $g(L, K)$ is a *convex function* due to the concavity of pY. Hence, the reliance on Kuhn-Tucker's original article by Averch-Johnson and Klevorick does not seem to be justified. However, as the Arrow-Hurwicz-Uzawa study has shown [2],[7] the Kuhn-Tucker-Lagrange conditions are relevant to this problem.[8] The Kuhn-Tucker-Lagrange conditions for this problem are:

a) $(1 - \lambda^*)(p + p'F)F_L \leqq (1 - \lambda^*)w$

b) $(1 - \lambda^*)(p + p'F)F_L L^* = (1 - \lambda^*)wL^*$

c) $(1 - \lambda^*)(p + p'F)F_K \leqq r - \lambda^* s$

d) $(1 - \lambda^*)(p + p'F)F_K K^* = (r - \lambda^* s)K^*$

e) $\lambda^*(pF(L, K) - wL^* - sK^*) = 0$

f) $\lambda^* \geqq 0.$

[3] Note this implies the concavity of the profit function which is defined in (3). Klevorick [5] assumes the concavity of the profit function, and this implies the concavity of the revenue function.

[4] Following the convention adopted by Averch-Johnson and Klevorick, depreciation is assumed to be equal to zero. Dropping the assumptions would not essentially alter the arguments.

[5] Westfield [8] formulated his problem to maximize the present value of the future net income stream. As he pointed out, this can be reduced to simple profit maximization.

[6] For the necessary conditions when g is concave, the Kuhn-Tucker constraint qualification can be satisfied if we assume that there exists $K'\geq 0$, $L'\geq 0$ such that $g(L', K')>0$ (Slater's condition). Averch-Johnson never refer to such conditions in stating their "Kuhn-Tucker" necessary conditions (8-1)-(8-6). (See [3], p. 1055.) They also claim that the concavity of the revenue function would guarantee the sufficiency, and in doing this, they have not referred to the shape of the constraint function. This is a bit unsatisfactory. See [3], fn. 3, p. 1055. See also footnote 8 of the present paper. The following quotation from Klevorick may be of interest to the reader. "... Averch and Johnson make use of the Kuhn-Tucker conditions for a constrained maximum. These conditions are sufficient as well as necessary under our assumption of concavity of profit function" [5], p. 479.

[7] Arrow-Hurwicz-Uzawa [2], especially Theorem 3, corollary 1. In other words the convexity of $g(L, K)$ would guarantee the necessity of the Kuhn-Tucker Lagrange conditions for a constraint maximum. Note also that $[g_L, g_K]\neq 0$ could also guarantee the necessity of these conditions for the constraint maximum. See Arrow-Hurwicz-Uzawa [2], corollary 6 of Theorem 3.

[8] But note also that only the case in which the constraint is *active* will be considered. Hence, to establish the necessary condition for the constrained maximum, the jargon of the Kuhn-Tucker theory, which was utilized by Averch-Johnson and Klevorick is not really needed. The classical Lagrangian method is more relevant to the discussion. For a good short discussion of the classical Lagrangian method, see Bliss [4], section 76, Theorem 76.3. The rank condition is obviously satisfied in our case. We may also note that, in general, the concavity of the profit function alone would not establish a sufficient condition for a maximum. The convexity of the constraint set is needed (as well as the concavity, or at least the quasi-concavity, of the maximum function) to establish a sufficiency condition. There is some difficulty in this connection due to the convexity (and not concavity) of the constraint function $g(L, K)$. This difficulty would still remain even under the assumption that the constraint is always effective. In the case of an effective constraint, λ can either be negative or positive. If λ is negative, than our Φ in (5) is obviously concave (hence quasi-concave) so that Bliss' condition in his Theorem 76.3 is satisfied. Again these points were never mentioned by Averch-Johnson and Klevorick.

Assuming that $L^*>0$, $K^*>0$ and $\lambda^*>0$ (i.e., the constraint (4) is *active* at (L^*, K^*),[9] the above conditions can be written:

$$(6) \qquad (p + p'F)F_L = w$$

$$(7) \qquad (1 - \lambda^*)(p + p'F)F_K = r - \lambda^* s$$

$$(8) \qquad pF - wL^* - sK^* = 0,$$

$$(9) \qquad \lambda^* > 0.$$

Notice that the three equations (6), (7), and (8) will determine the values of L^*, K^* and λ^*.[10]

If there is no *fair rate of return* constraint, or the constraint is not *active*, equations (6) and (7) become:

$$(10) \qquad (p + p'F)F_L = w$$

$$(11) \qquad (p + p'F)F_K = r$$

Since $(p+p'F)=d(pF)/dY$, (10) and (11) are nothing but the familiar rule which equates the marginal revenue product of each factor to its price. Assume that there exist finite optimal values L_0 and K_0 for this nonconstrained maximization problem. Then the nonconstrained rate of return on capital for the nonconstrained case can be computed by:

$$(12) \qquad r_0 = \frac{pF(L_0, K_0) - wL_0}{K_0}$$

If $r_0 \leqq s$, then the imposition of a fair rate of return constraint would not affect the solution of (6), (7), and (8); hence it must be assumed that $r_0>s$ or that the constraint is active to discuss the Averch-Johnson case. These authors claim $\lambda^*<1$. To see the effect of this, we rewrite equation (7) as:

$$(13) \quad (p + p'F)F_K = r - \frac{s - r}{1 - \lambda^*}\lambda^* = \bar{r}$$

Under the assumption of $s>r$ and $\lambda^*<1$, $\bar{r}<r$. The concavity of the revenue function, $G \equiv pF(L, K)$, implies that its Hessian matrix is negative semidefinite. Hence, in particular, we have:

$$(14) \quad \frac{\partial^2 G}{\partial K^2} = \frac{\partial}{\partial K}\left[(p + p'F)F_K\right] \leqq 0$$

In other words, $\bar{r}<r$ implies that the amount of capital used with the regulatory constraint (K^*) is not less than the amount used without a constraint (K_0), provided that the marginal-revenue-product-of-capital *MRPK* curve does not shift. If G is assumed to be strictly concave, then $G_{KK}<0$; hence $K^*>K_0$.[11] This seems to be a more or less desired conclusion by Averch-Johnson [3 p. 1056] and Klevorick. Notice that the concavity of the revenue function G plays an important role in the above consideration but this is not mentioned in Averch-Johnson and Klevorick.

Now we should point out the most crucial assumption in the above argument, namely that the function which represents the marginal revenue product of capital *MRPK* does not shift. This is not true in general, and this causes a serious criticism against the main arguments of Averch and Johnson. It involves an error of confusing the movement along the curve and the shift of the curve. With an introduction of the fair rate of return regulation, the firm will adjust the amount of capital and labor inputs so that it would be optimal under regulation. If the amount of labor em-

[9] These are assumptions made by Averch-Johnson and others. There is no a priori reason why this is so. For example, if the demand curve is of a constant elasticity type ($p=AY^{-1/\eta}$ where η is the elasticity) and if $\eta<1$, then the marginal revenue curve function $(p+p'F)$ is always negative, and thus $L^*=0$.

[10] We assume that the proper conditions (such as the nonvanishing Jacobians) should be true.

[11] $G_{KK}<0$ says that the marginal revenue product of capital is a decreasing function of capital used. This is not at all a weak assumption. The marginal value product of each factor is the product of marginal revenue and the marginal physical product of factor, and we can note easily: 1. the marginal physical product of capital is not necessarily a decreasing function of the amount of capital used if the production function exhibits an increasing returns to scale for example (e.g. $Y=L^{0.5}K^2$); and 2. marginal revenue is not necessarily a decreasing function of output even if the demand function is negatively sloped (e.g. $p=Y^{-2}+1$).

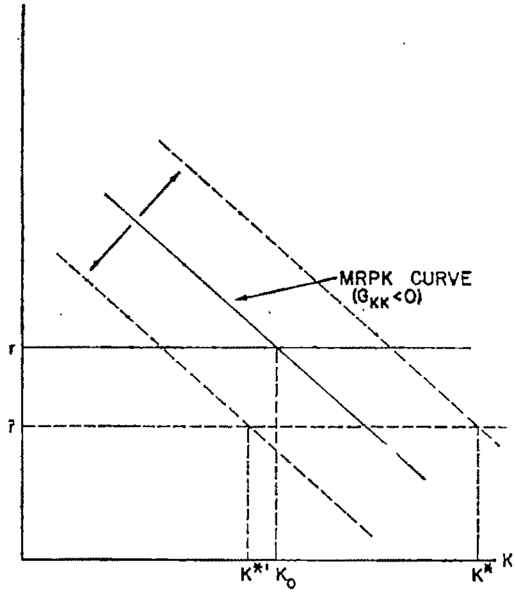


FIGURE 1

ployed should be altered, this would cause a shift of the *MRPK* curve, for the marginal product of capital at each level of capital used is, in general, a function of the labor employed. As it can be illustrated from Figure 1, if the *MRPK* curve shifts downwards too much, then, even under the strict concavity of G, we cannot conclude that $\bar{r}<r$ implies $K^*>K_0$, and there is no a priori reason to believe that *MRPK* does not shift downwards too much. In Figure 1, $K^{*\prime}$ represents the case in which the MRPK curve shifts downwards too much so that $K^{*\prime}$ can be less than K_0. In short, until we can specify in which direction the *MRPK* would shift and how much it would shift, we cannot tell whether the introduction of a *fair rate of return* would increase the amount of capital used by the firm.[12]

[12] We may note that we can conceive a case in which the *MRPK* curve does not shift downwards at all due to some institutional reasons. For example, if the labor union of the firm strongly opposes any reduction of labor employment, then the *MRPK* curve would not shift downwards. Hence, under the assumption of strict concavity, it follows that $\bar{r}<r$ indeed implies $K^*>K_0$, without any further consideration.

Another important question at this point is whether λ^* is indeed less than one. The proof by Averch-Johnson roughly goes as follows: Since $s>r$, λ^* can not be equal to one, due to equation (7). When $s=r_0$ (the unconstrained rate of return), $\lambda^*=0$. Because the continuity of λ^* with respect to s, λ^* should always be less than one.

Although this proof is rather ingenious, it crucially depends on the continuity of λ^* in s. Hence, it may be worthwhile to specify the proof of the continuity of λ^*, or at least to specify the conditions for its continuity. Averch and Johnson do not provide any such proof or considerations, and the continuity of λ^* is not intuitively obvious, for, in general, the value of the Lagrangian multiplier may jump from zero to some non-zero value as the constraint moves from the *non-active* stage to the *active* stage. In this example λ^* may appear to be discontinuous at the point of $s=r_0$, where the constraint (4) moves from the *non-active* stage to the *active* stage. We shall find out the conditions under which λ^* is continuous in this example. First, note that we can explicitly obtain the value of λ^* from equation (7), assuming $G_K-s\neq 0$.[13] In other words:

$$\lambda^* = \frac{G_K - r}{G_K - s} \tag{17}$$

If $G_K-s=0$, then due to (7), $(1-\lambda^*)s=r-\lambda^*s$ so that $s=r$, which contradicts the assumption of $s>r$. Hence, the equation (17) would be valid for all cases under the present consideration.[14] Now if we

[13] Recall the notations: $G(L, K)\equiv pF(L, K)$ and $G_K\equiv\partial G/\partial K$. Hence $G_K=(p+p'F)F_K$.

[14] Note that this would imply a certain restriction for the class of functions which is eligible for our analysis. For example, if the revenue function G is homogeneous of degree one, then $G_K=s$ due to (6) and (8), so that such a revenue function is not eligible. For a certain class of functions, equation (7) with $s>r$, and the assumption of "active" constraint (i.e., equation (8)) are not consistent with each other.

assume that G_K is a continuous function of L^* and K^* and that L^* and K^* are continuous functions of s, the right hand side of equation (17) is continuous with respect to s, so that λ^* is a continuous function of s. In other words the continuity of λ^* depends on the continuity of L^* and K^* with respect to s.

Averch-Johnson [3] and Klevorick [5] argue that the allocative efficiency of inputs should be measured by the ratio of F_K/F_L to r/w. This ratio would be less than one as long as $\lambda^*<1$, as they claim. However, it is the present contention of the author that this criterion does not have convincing support as a measure of efficiency, for such an allocative principle would be valid under the competitive situation only, and it will not be generally valid under other situations according to the second best Theorem.

II *Applications*

The considerations noted may cause the reader to think that we can not easily specify whether the value of K^* (the constrained case) is larger than the value of K_0 (the nonconstrained case). However, this is not the case. To see this, we first suppose that the regulatory agency initially sets the value of s equal to r_0 the rate of return on capital in the non-constrained case, and then the agency reduces this rate s below r_0. Then the corresponding optimal values of K and L will vary. The problem is to find the direction of changes of K and L for a change in s and this is an elementary exercise of comparative statics. Rewrite new equations (6), (13), and (8) as follows (recalling $G(L, K)\equiv pF(L, K)$):

$$G_L = w \tag{18}$$

$$G_K = r - \frac{s-r}{1-\lambda^*}\lambda^* \tag{19}$$

$$G - wL^* - sK^* = 0 \tag{20}$$

As remarked previously, these three equations are sufficient to specify values of L^*, K^* and λ^*. Note now that equations (18) and (20) form a subsystem within which the values of L^* and K^* are independent of λ^*. Assuming the left-hand differentiability of L^* and K^* with respect to s at the point $s=r_0$, differentiating (18) and (20) with respect to s[15] yields the following equations:

$$G_{LL}\frac{dL^*}{ds} + G_{LK}\frac{dK^*}{ds} = 0 \tag{21}$$

$$\frac{dL^*}{ds}(G_L - w) + \frac{dK^*}{ds}(G_K - s) = K^* \tag{22}$$

Note that $s=r_0$ initially and in view of (18) and (19), equation (22) can be rewritten as:[16,17]

$$\frac{dK^*}{ds} = \frac{-K^*}{r_0 - r} < 0. \tag{23}$$

Recalling that $r_0>r$, the introduction of an *active* constraint of the *fair rate of return* type would indeed increase the amount of capital invested by the firm. This is the result claimed by Averch-Johnson. Note that the crucial equation in the above logic is equation (19) (or (8)), which essentially states that the constraint is "active."

The change in the quantity of labor can now be obtained by inserting (23) into (21).

$$\frac{dL^*}{ds} = \frac{G_{LK}}{G_{LL}}\frac{K^*}{r_0 - r} \tag{24}$$

Assume $G_{LL}<0$ and $G_{LK}>0$, ($G_{LL}<0$ is true under the strict concavity of G).[18] Then

[15] It is not assumed that λ^* is differentiable with respect to s.

[16] Note that $G_K=r$ initially since $\lambda^*=0$ (see equation (19)).

[17] It may be of interest to investigate what would happen to the total profit. Differentiating (3) yields: $d\pi/ds=(G_L-w)\ dL^*/ds+(G_K-r)dK^*/ds$. Hence, in view of (18) and (19) with $\lambda^*=0$, we may conclude that $d\pi/ds=0$ at $s=r_0$.

[18] The concavity of G requires $G_{LL}\leq 0$. Its strict concavity requires $G_{LL}<0$.

$dL^*/ds<0$, i.e., the employment of labor will increase under these assumptions.

The value of dK^*/ds at the point of $s=r_0$ was obtained in (23). Yet, in view of (19) and (23), the value of dK^*/ds for all the values of s such that $r<s\leqq r_0$ can be obtained by assuming the differentiability of L^* and K^* with respect to s in this range.[19]

$$(25)\quad \frac{dK^*}{ds}=\frac{K^*}{r-s}(1-\lambda^*)=\frac{K^*}{G_K-s}$$

When $s=r_0$, $\lambda^*=0$ as noted above, and thus $dK^*/ds=K^*/(r-r_0)<0$. Equation (25) implies that $dK^*/ds\neq 0$ for any finite value of G_K. Assume that G_K is finite for all $L>0$ and $K>0$. Hence, assuming dK^*/ds is a continuous function of s, and $r<s\leqq r_0$, it follows $dK^*/ds<0$ always. In other words, $K^*>K_0$ for all the values of s such that $r<s\leqq r_0$, i.e., $K^*>K_0$ is true not only for the small displacement of ds from r_0 but also for all the displacement of s from r_0 as long as $r<s\leqq r_0$.

Finally a comment on the assumption of the concave revenue function is in order. Although this is a very convenient assumption to make in solving the above maximization problems, sometimes it is not necessarily an easy assumption to justify. For example, a linear homogeneous concave production function does not guarantee the concavity of the revenue function G. If we assume the forms of p and Y as $p=AY^{-1/\eta}$, $\eta>0$, $A>0$ and $Y=L^\alpha K^\beta$, α, $\beta>0$, then the revenue function can be written as:

$$(26)\quad G(L,K)=AL^{\epsilon\alpha}K^{\epsilon\beta} \quad \text{where } \epsilon=1-1/\eta,\quad 0<\epsilon<1.$$

This function is concave if $\epsilon\alpha+\epsilon\beta<1$. If $\epsilon\alpha+\epsilon\beta>1$, then the function is quasi-concave but is no longer concave. If $G(L,K)$ is quasi-concave but not concave, then the profit function defined in (3) is no longer quasi-concave, and the line of thought developed in Arrow-Enthoven [6] can not be extended for application to this case.

[19] The change of profit again can be computed for the value of s such that $r<s\leqq r_0$. I.e., $d\pi/ds=\lambda^*K^*\geqq 0$ (see the previous footnote and equations (17), (18), and (25)). In other words, the profit decreases as the fair rate of return decreases as long as $\lambda^*>0$. Since $K^*>K_0$, the profit rate $(pY-wL-rK)/K$ also decreases.

References

1. K. J. Arrow, and A. C. Enthoven, "Quasi-Concave Programming," *Econometrica*, Oct. 1961, *29*, 779–800.
2. ——— L. Hurwicz, and H. Uzawa, "Constraint Qualifications in Maximization Problems," *Naval Research Logistics Quarterly*, Jan. 1961, *8*.
3. H. Averch, and L. O. Johnson, "Behavior of the Firm Under Regulatory Constraint," *Am. Ec. Rev.*, Dec. 1962, *52*, 1052–69.
4. G. M. Bliss, *Lectures on the Calculus of Variations*, Chicago 1946.
5. A. K. Klevorick, "Graduated Fair Return: A Regulatory Proposal," *Am. Ec. Rev.*, June 1966, *56*, 477–84.
6. H. W. Kuhn, and A. W. Tucker, "Nonlinear Programming," In J. Neyman, Ed., *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley and Los Angeles 1951.
7. S. H. Wellisz, "Regulation of Natural Gas Pipeline Companies: An Economic Analysis," *Jour. Pol. Econ.*, Feb. 1963, *71*, 30–43.
8. F. M. Westfield, "Regulation and Conspiracy," *Am. Ec. Rev.*, June 1965, *55*, 424–43.

Tariffs, Intermediate Goods, and Domestic Protection

By Roy J. Ruffin*

The recognition of imported intermediate goods has resulted in a popular new measure of domestic protection that takes into account the tariffs on material inputs as well as the tariff on final output.[1] This new measure, called the *effective protective tariff* rate or simply *effective tariff*, turns out to be zero if the tariffs on intermediate goods wipe out the apparent advantages of the tariff on final output. With the aid of the effective tariff concept, investigators have discovered many interesting facts about the U. S. tariff structure: for example, the effective tariff on Ingots and Steel Forms is ten times as high as the nominal rate; the effective tariff on Paper and Paper Products is almost zero; and the rate on Agricultural Machinery is negative.[2] This new theory of tariffs, as G. Basevi [3] and W. M. Corden [5] refer to the effective tariff, appears to have increased our empirical knowledge of the extent of protection.

Yet, the theoretical basis for the effective tariff rate has not been made clear. All the studies use an input-output model that is, as Harry Johnson [7] has admitted, "too restrictive for most analytical purposes." And without a simple theory of trade in intermediate goods to guide them, the authors of these studies can hardly be blamed for falling back on familiar tools.

The purpose of this paper is to develop a simple approach to the pure theory of trade in intermediate goods that can be used to answer some analytical questions raised by the concept and measurement of the effective tariff rate. It is hoped that the approach will be useful for other purposes as well.

At the outset, we should note that extravagant claims have not been made for the effective tariff rate concept. The various writers have claimed that the new measure is better than nominal tariff rates and this claim will not be disputed here. But any measure that is of potential use to policy makers must be examined to determine its positive and normative content. I think it is fair to say that virtually no attempt has been made to do this. Indeed, the effective tariff rate concept has not even been defined in a general equilibrium context and, as a result, the current measure does not always measure what it is supposed to.

Sections I, II, and III contain an extensive discussion of a simple small country model in which the effective tariff concept has a solid theoretical foundation, but in which the current *measure*—called the Barber-Johnson-Balassa formula—is valid only for a special case. The crucial assumption of this model is that imported intermediate goods cannot be produced domestically. In Section IV, it is shown that when this crucial assumption is relaxed the effective tariff concept loses certain desirable properties. Finally, Section V summarizes the argument and draws

* The author is assistant professor of economics at Washington State University. He is indebted to his colleague Thomas A. Lowinger for valuable discussion of a previous version of this paper and to the Bureau of Economic and Business Research of Washington State University for financial support. He also wishes to thank Millard Hastay of Washington State for suggesting improvements in the manuscript.

[1] This new measure was discovered by Barber [2] and developed by Balassa [1], Basevi [3], Corden [4] [5], and Johnson [7].

[2] These estimates are taken from Balassa [1].

conclusions from the findings.

Apart from the model itself, the main tool used for evaluating the effective tariff concept is the theory of second best. For our problem the theory of second best renders valuable services, since by its use the argument is greatly simplified. Furthermore, what we need of this theory is minimal and can be proved in the text. It is applicable because, in a small country world, a first best optimum cannot be achieved when tariffs are imposed on the imports.

Throughout the paper we assume one exported final good, one imported final good, and one imported intermediate good. This simplifies the discussion and does not sacrifice any matters of principle; it also allows geometrical demonstrations.

I. *A Simple Model*

Consider a small country producing two consumer goods. Initially, it will be supposed that one of these goods requires, in fixed proportions, an intermediate good that cannot be produced at home. The problem of income distribution is ignored by postulating a collective utility function. Assumptions A.1 through A.5 describe the model in detail:

- A.1. There are three goods. X_1 and X_2 are final goods produced domestically and X_3 is an imported intermediate good.
- A.2. Preferences are represented by a collective utility function, $U = U(C_1, C_2)$, where C_i is the consumption of good i. The collective indifference curves are convex to the origin.
- A.3. Each unit of X_2 requires a units of X_3, i.e., $X_3 = aX_2$.
- A.4. The gross transformation curve, which defines production possibilities on the assumption the requisite amount of X_3 has been obtained, is concave to the origin and is given by $T(X_1, X_2) = 0$.
- A.5. The prices P_1, P_2, and P_3 are fixed on world markets.

Later, we shall replace assumption A.3 by:

- A.3′. Each unit of $X_i (i=1, 2)$ requires a_i units of X_3.

A few brief remarks on these assumptions may be worthwhile. Assumption A.2 on preferences has a long and, perhaps, not so honorable tradition in trade theory; we shall let the reader decide on its worth. The assumption, A.3, that the intermediate good is used in fixed proportions is a simplifying assumption, which is shared by the empirical studies on the effective tariff rate. A similar comment applies to the small country assumption, A.5. The assumption, A.4, that X_3 does not enter the transformation function means, of course, that it cannot be produced. It is crucial and will be relaxed in Section IV. In Section III we replace assumption A.3 with A.3′ and the results are dramatic.

Consider now the purely technical problem of maximizing the utility function subject to the constraints. What are the constraints? For any pair (X_1, X_2) satisfying the gross transformation function,

$$T(X_1, X_2) = 0, \tag{1}$$

the economy must pay for the required X_3 by satisfying the budget constraint,

$$P_1X_1 + P_2X_2 - P_3X_3 = P_1C_1 + P_2C_2.$$

Using A.3, this constraint may be written in the more convenient form:

$$P_1X_1 + P_2'X_2 = P_1C_1 + P_2C_2 \tag{2}$$

where $P_2' = P_2 - aP_3$. P_2' is the contribution of a marginal unit of X_2 to private and social income and may be interpreted as the *net* price or the *value added* of X_2.

If $U = U(C_1, C_2)$ is maximized subject

to constraints (1) and (2), the necessary conditions for a maximum are:

$$U_1/U_2 = P_1/P_2 \tag{3}$$

$$T_1/T_2 = P_1/P_2' \tag{4}$$

Thus, since the static stability conditions are met under A.1 through A.5, a central planner could turn his job over to maximizing consumers (more precisely, the consumer representative) and business firms.

A geometric picture will prove useful. Businessmen maximize the function $P_1X_1 + P_2'X_2$ on $T(X_1, X_2) = 0$. Given the optimal decision on X_1 and X_2, the budget constraint (2) for consumers is determined and consumers maximize in the usual way. In Figure 1, TT' is the transformation curve; RR' is the business budget constraint, with slope $-P_2'/P_1$; Rr is the consumer budget constraint, with slope $-P_2/P_1$; and u is the highest attainable indifference curve. Note that in the diagram at the optimal position X_1 is exported, X_2 imported. But the indifference curve could be tangent anywhere on Rr, giving any desired combination. For example, if the consumption point were southwest of point E, both X_1 and X_2 would be exported. The budget constraint insures that the total value of exports equals the total value of imports, including the intermediate good. And it is an easy exercise to show that SE measures the imports of X_3 in terms of X_2.[3]

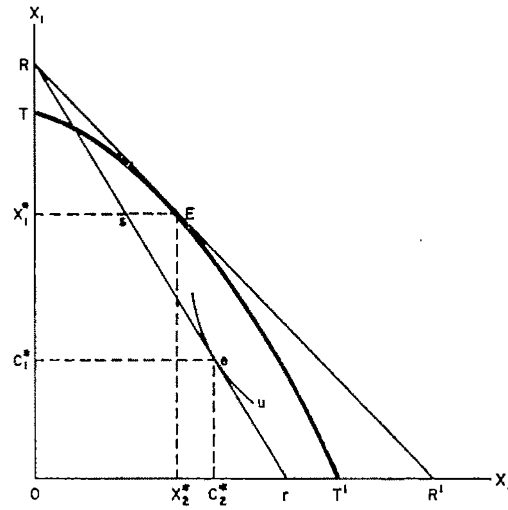


FIGURE 1

the attainment of the necessary marginal conditions, (3) and (4), for a maximum. The right hand sides of these equations measure marginal social costs (with trade) of transforming one good into another. Tariffs destroy (3) and (4) by inducing consumers and business firms to equate their marginal rates of substitution or transformation to marginal private costs, which with tariffs diverge from marginal social costs. Hence, for our model as well as the traditional model, the optimal tariff on each good is zero when prices are fixed on world markets. The question then arises: when an arbitrary positive tariff is imposed on X_1 or X_2, what is the second best tariff on the intermediate good, X_3? The answer to this question is helpful in evaluating the concept of effective tariffs.

We first prove that if one of the marginal equalities (3) and (4) is not satisfied, the other represents a condition for a second best optimum.

II. *Some Effects of Tariffs*

As a prelude to discussing the effective tariff concept, we must state and prove an interesting theorem about tariffs in the above model. Two more such theorems are proved in the next section.

Consider, then, the effects of a tariff on

Fundamental Lemma. If $U_1/U_2 = kP_1/P_2$ $(k \neq 1)$, then under assumptions A.1 through A.5, equation (4) is a necessary condition for a second best optimum; and if $T_1/T_2 = mP_1/P_2'(m \neq 1)$, then under assumptions A.1 through A.5,

[3] In Figure 1, $OR = X^*_1 + (P'_2/P_1)X^*_2$. Thus, the distance between point R and point X_1^* is $(P'_2/P_1)X^*_2$. Since the (absolute) slope of Rr is P_2/P_1, the distance between point X_1^* and point S must be $(P'_2/P_2)X_2^*$. It follows that $SE = aP_3X_2^*/P_2$.

equation (3) is a necessary condition for a second best optimum.

Proof. To establish the first part of the lemma, maximize $U=U(C_1, C_2)$ subject to (1), (2), and $U_1/U_2=kP_1/P_2$. This results in:

$$U_i-\lambda_1 P_i+\lambda_2(U_2U_{1i}-U_1U_{2i})/U_2^2=0, \quad i=1, 2$$

$$T_1/T_2=P_1/P_2' \tag{4}$$

which proves that (4) is necessary for a maximum. For the second part, maximize $U=U(C_1. C_2)$ subject to (1), (2), and $T_1/T_2=mP_1/P_2'$. This results in:

$$U_1/U_2=P_1/P_2 \tag{3}$$

$$\lambda_0 T_i+\lambda_1 P_i'+\lambda_3(T_2T_{1i}-T_1T_{2i})/T_2^2=0, \quad i=1, 2$$

which proves that (3) is necessary for a maximum.

There is nothing surprising about this lemma. It is a special case of a more general theorem proved by O. A. Davis and A. B. Whinston;[4] and it holds for many economic models, including the traditional trade model. Its significance arises from the fact that with our simple model there exists a tariff structure such that one marginal equality will be satisfied while the other is not. In our model in Section IV, if tariffs interfere with consumer choice then they must interfere with producer choice as well; thus, in that section the lemma will not be so useful, although it will still be true.

Now let t_i denote the tariff on the ith good. The Fundamental Lemma makes it easy to prove the following theorem:

Theorem 1. Given $t_1=0$ and any $t_2>0$, then under assumptions A.1 through A.5 the second best tariff, t_3^*, on X_3 is

$$t_3^* = t_2P_2/aP_3. \tag{5}$$

[4] The interested reader is referred to Davis and Whinston [6] and the important comments by McManus [9].

Proof. Assume that $t_1=0$, $t_2>0$, and t_3 is any number. Under free competition, it is clear that the following marginal conditions must be satisfied:

$$U_1/U_2 = P_1/P_2(1+t_2) \tag{6}$$

$$T_1/T_2 = P_1/[P_2(1+t_2) - aP_3(1+t_3)]. \tag{7}$$

Since under our assumptions we can only vary t_3, to reach a second best optimum the right hand side of (7) must be the same as the right hand side of (4). Thus, solving the equation,

$$P_2(1+t_2) - aP_3(1+t_3^*) = P_2 - aP_3,$$

yields equation (5).

From Theorem 1, it is easy to see that the second best tariff on the intermediate good is positive and larger than the tariff on X_2. Before applying this theorem to the concept of effective tariffs, an independent geometric explanation will be useful.

In Figure 2, RR' and Rr are the pre-tariff budget lines for business firms and consumers, respectively. Business firms choose point E on TT' and consumers point e on indifference curve u. When a positive tariff is imposed on X_2, without

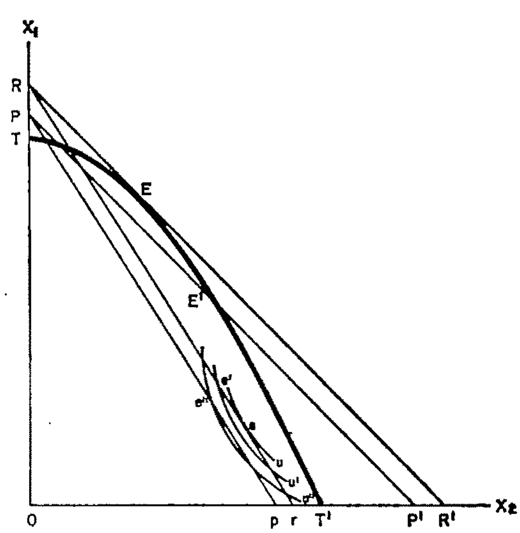


FIGURE 2

any tariff on X_3, the new production point will be at a point such as E', where the output of X_2 is larger than at E. Since PP' is the new budget line for business firms, Pp is the new budget line for consumers. Consumers will not move to an indifference curve tangent to Pp, as the slope of Pp measures marginal *social* costs and not marginal *private* costs. Thus, they will choose some such point as e'' on indifference curve u''. If now a positive tariff is imposed on the intermediate good, X_3, business firms will reduce X_2 production and expand X_1 production, moving the production point back towards E. The second best tariff on X_3 induces business firms to choose point E, the pre-tariff production point; and it will be larger than the tariff on X_2, since the cost of X_3 is only a fraction of the total cost of producing X_2 and the net protection must be zero. Hence, the budget line for consumers shifts upward from Pp to Rr and consumers will choose a point such as e' on u'. This is clearly not as good as the pre-tariff equilibrium, e on u, but it is better than e''. If the tariff on the intermediate good is increased beyond the second best level, the production point moves to the northwest of E and the budget line falls below Rr. Finally, note that as t_3 increases from 0 to t_3^*, the second best tariff, welfare steadily increases.

III. *The Effective Tariff: Intermediate Goods Not Produced Domestically*

Theorem 1 will now be used to investigate the properties of the effective tariff rate. But first three definitions are needed.

D.1. Let P_j' be the domestic value added or net price of the jth good; and let the kth final good be free of subsidies and tariffs. Then a measure, t_i', is said to be the effective tariff rate (relative to k) on industry i if t_i' is positive, zero, or negative, if and only if, the marginal rate of transformation, T_i/T_k, exceeds, equals, or falls short of P_i'/P_k'.[5]

D.2. If when the effective tariff rate on the imported final good is zero and the tariff on the final good is nonzero, the production point of the economy is the same as the production point under free trade, then the effective tariff is said to be *unbiased.*

D.3. If when the effective tariff on the imported final good is zero and the tariff on the final good is nonzero, a second best optimum is achieved, then the effective tariff is said to be *efficient.*

Note that D.1 is framed in such a way that the concept of the effective traiff is distinguished from the derived measure in any particular model. D.2 and D.3 are desirable positive and normative properties of the effective tariff rate.

Now let us examine the Barber-Johnson-Balassa formula for measuring effective tariffs. Their measure is:

$$t_i'' = (t_i - \sum_j a_{ij}t_j)/v_i \tag{8}$$

where a_{ij} = the value of input j used in a dollar's worth of good i and v_i = the value added in producing a dollar's worth of i. Applying this formula to our model, we find that:

$$t_2'' = (P_2t_2 - aP_3t_3)/P_2'. \tag{9}$$

[5] The motivation for this definition can be seen by referring to Figure 1 or 2. When the effective tariff on X_2 is positive, the production point will be to the southeast of point E. At such a point, the marginal rate of transformation of X_1 into a marginal unit of X_2, T_2/T_1, is greater than P_2'/P_1. This assumes that there are no governmental or monopolistic impediments, aside from the tariffs themselves, to the competitive adjustment of markets.

It should also be noted that the definition is stated for an n-commodity world to eliminate any possible confusion that might result from a more limited definition. It can be shown that when the kth and the jth final goods do not use the imported intermediate goods, the effective tariff on good i relative to k equals the effective tariff on good i relative to j.

It is easy to see that (9) satisfies our definition (D.1) and that the effective tariff rate is both unbiased and efficient. When $t_2'' = 0$, $t_3 = t_3^*$ (the second best tariff) and production will be at point E. When $t_2'' > 0$, the production point will be to the southeast of point E; and when $t_2'' < 0$, the production point will be to the northwest of point E.

Theorem 1 thus provides us with a neat theoretical justification of the Barber-Johnson-Balassa measure of the effective tariff rate. But our simple model can also be used to show a serious limitation to this measure. For practical purposes, Theorem 1 depends on the assumption that the good using the imported intermediate good is imported as well, for tariffs are not usually imposed on exports. Many countries, however, have a comparative advantage in the processing of the raw materials and semi-finished manufactures of other countries. This leads us to the question: what is the second best tariff on X_3 when X_1 is imported?

Theorem 2. Given $t_2 = 0$ and $t_1 > 0$, then under assumptions A.1 through A.5 the second best tariff on X_3, t_3^*, is:

$$t_3^* = -\, t_1 P_2'/aP_3. \tag{10}$$

Proof. The proof follows that of Theorem 1. Under free competition,

$$U_1/U_2 = P_1(1 + t_1)/P_2 \tag{11}$$

$$T_1/T_2 = P_1(1 + t_1) \,/[P_2 - aP_3(1 + t_3)] \tag{12}$$

As before, we solve for the t_3^* which equates the right hand side of (12) to the right hand side of (4). This gives us (10).

When the tariff is imposed on the good not requiring the intermediate good, the second best tariff on X_3 is negative! Figure 2 may also be used to explain this theorem. If t_1 is positive and t_3 zero, business firms will choose a production point northwest of point E. The second best optimum, however, requires that the economy be at point E. To move back to point E, under the assumptions of the theorem we can only impose an appropriate negative tariff on X_3.

Let us now replace assumption A.3 with assumption A.3′. Recall that Theorem 1 stated that when X_2 is imported subject to tariff and is the only good using the imported intermediate good, the second best tariff on X_3 is positive and larger than the tariff on X_2. Theorems 1 and 2 suggest that if X_3 were used in both X_1 and X_2, the second best tariff on X_3 (when X_2 is imported subject to tariff) may be of any sign or magnitude. The effective tariff, as currently measured, will then miss the mark completely on both protection and welfare grounds; moreover, it will not satisfy our reasonable definition of the effective tariff rate.

The following theorem can be used for evaluating the effective tariff rate under these new circumstances.

Theorem 3. Given $t_1 = 0$ and $t_2 > 0$, then under assumptions A.1, A.2, A.3′, A.4, and A.5, the second best tariff on X_3 is:

$$t_3^* = -\, t_2 P_2 P_1'/P_3(a_1 P_2 - a_2 P_1) \tag{13}$$

where $P_1' = P_1 - a_1 P_3$.

Proof. Under free competition,

$$U_1/U_2 = P_1/P_2(1 + t_2) \tag{14}$$

$$T_1/T_2 = \frac{P_1 - a_1 P_3(1 + t_3)}{P_2(1 + t_2) - a_2 P_3(1 + t_3)}\,. \tag{15}$$

As before, we solve for the t_3^* that equates the right side of (16) to the appropriate marginal condition, which in this case is P_1'/P_2'.[6]

[6] With assumption A.3′ replacing A.3, the maximization problem is as follows: maximize $U = U(C_1, C_2)$ subject to $T(X_1, X_2) = 0$ and

$$P_1' X_1 + P_2' X_2 = P_1 C_1 + P_2 C_2.$$

Accordingly, the necessary conditions for a maximum are $U_1/U_2 = P_1/P_2$ and $T_1/T_2 = P_1'/P_2'$. When consump-

It is easy to check that the following measure of the effective tariff rate satisfies the definition and that the measure is both unbiased and efficient:

$$(16)\qquad t_2' = [t_2P_2P_1' + t_3P_3(a_1P_2 - a_2P_1)]/P_2'$$

The implications of (16) are best seen by rearranging the terms and setting $P_1=1$. Thus:

$$(17)\qquad t_2' = [(P_2t_2 - a_2P_3t_3) + a_1P_2P_3(t_3 - t_2)]/P_2'$$

Now compare our measure, equation (17), with the Barber-Johnson-Balassa measure for this particular model:

$$(18)\qquad t_2'' = (P_2t_2 - a_2P_3t_3)/P_2'.$$

Since tariffs on intermediate goods, t_3, are usually lower than the tariffs on final goods, t_2, it follows that $t_2' < t_2''$. The Barber-Johnson-Balassa formula overestimates the correct effective tariff rate![7]

In a model with only one imported final good (and hence only one effective tariff rate), this is perhaps not too serious. But when there are several effective tariff rates, the overestimates on each effective tariff rate can drastically change the relative sizes of the various rates and their rankings.

So far our model has been quite favorable to the effective tariff concept, although it has not been so kind to the current measure. We now turn to a model that is not so kind to the concept itself.

IV. *The Effective Tariff: Intermediate Goods Produced Domestically*

What happens to the effective tariff concept when the imported intermediate good is produced domestically? In this section it is shown that under these conditions the effective tariff rate is both biased and inefficient. Bias means that an industry may still be protected or injured as compared to free trade output even when the effective tariff rate is zero. This gives, in part, formal expression to Viner's view that protection cannot be precisely measured.[8] Inefficiency means that a higher effective tariff rate may be better than a lower one.

For ease of exposition, we will go back to assumption A.3, that the intermediate good is used only in X_2. Thus the effective tariff rate will be measured by the Barber-Johnson-Balassa formula.

With the work done in previous sections, the new model is easy to state. This model consists of assumptions A.1, A.2, A.3, A.5, the gross transformation surface is concave to the origin and is given by $T(X_1, X_2, X_3)=0$. The budget constraint is now,

$$P_1X_1 + P_2X_2 + P_3X_3 = P_1C_1 + P_2C_2 + P_3\overline{X}_3,$$

where $\overline{X}_3 = aX_2$. This may be rewritten as:

$$(19)\qquad P_1X_1 + P_2'X_2 + P_3X_3 = P_1C_1 + P_2C_2.$$

The new necessary conditions for maximum utility are now obtained by maximizing $U=U(C_1, C_2)$, subject to $T(X_1, X_2, X_3)=0$ and (19). They are

$$(20)\qquad U_1/U_2 = P_1/P_2$$

$$(21)\qquad T_1/T_2 = P_1/P_2'$$

$$(22)\qquad T_1/T_3 = P_1/P_3$$

Thus, now that X_3 is produced domestically, a maximum requires the marginal rate of transformation between one of the goods and X_3 to be equal to their price ratio. This has important implications.

tion decisions are distorted by tariffs, the latter condition is necessary for a second best optimum.

[7] For completely different reasons, W. M. Corden [5, pp. 225 and 235] and Ronald I. McKinnon [8, p. 612] have reached a similar sounding conclusion.

[8] For the relevant quote from Viner, see Balassa [1, p. 573].

Suppose that one imposes a tariff on X_2 and a tariff on X_3, denoted by $\bar{t}_3$, such that

$$P_2' = P_2(1 + t_2) - aP_3(1 + \bar{t}_3),$$

that is, the effective tariff on X_2 is zero. Under free competition, it is clear that

$$(23) \qquad U_1/U_2 = P_1/P_2(1 + t_2)$$

$$(24) \qquad T_1/T_2 = P_1/P_2'$$

$$(25) \qquad T_1/T_3 = P_1/P_3(1 + \bar{t}_3)$$

Now marginal equalities (21) and (22) are necessary for the production point to be at the free trade point. However, from (25) we see that (22) is not satisfied. It is immediate, therefore, that in this new model the effective tariff rate is biased.

With this tariff structure, one cannot say whether the output of X_2 will be greater than, equal to, or less than the free trade output. More X_3 will be produced at home; but this can be accomplished in several ways. The production of *both* X_1 and X_2 may be reduced, or that of one increased, and the other decreased. It all depends on the conditions of production.

The welfare implications are equally uncertain. If marginal equality (20) is not satisfied, then marginal equalities (21) and (22) must be satisfied for a second best optimum. Thus, suppose we start with t_2 positive and then impose a tariff on X_3 such that the effective tariff is zero. With just a tariff on X_2, only (22) is satisfied; with the effective tariff on X_2 zero, only (21) is satisfied. It is impossible to say which of these situations is better; for the zero effective tariff corrects one marginal condition and simultaneously destroys another. Hence, we cannot rule out the possibility that an extremely high effective tariff level is better than a zero effective tariff level.

V. *Conclusions*

When imported intermediate goods are not produced domestically, the effective tariff concept has a strong theoretical foundation. Properly measured, the effective tariff rate has implications for both protection (unbiasedness) and welfare (efficiency). The Barber-Johnson-Balassa formula, however, is correct only under the special assumption that exports do not use the imported intermediate good. Thus, the first task for future research would seem to be to modify existing estimates of effective tariff rates. Furthermore, as long as tariffs on intermediate goods are lower than on final goods, the existing estimates should be overestimates of the correct effective tariff rates.

Without an empirical investigation, it is (of course) impossible to say whether correct estimates of effective tariff rates would significantly change the relative rankings and industry rates of protection found in the published empirical studies. But these studies cannot be taken too seriously until revised or at least checked.

A caveat, however, must be issued. There is a formidable difficulty that stands in the way of a neat application of the effective tariff concept to a many commodity world. As is clear from our definition, the effective *protection* of one industry can only be measured relative to some base industry. This is as it should be, for comparative and not absolute advantages direct the allocation of resources. Unfortunately, it means that there is no unique measure of protection in any one industry. Each effective tariff rate depends on the base industry to which the industry in question is compared; and two different bases may lead to entirely different conclusions. The partial equilibrium approach of Barber, Balassa, Basevi, Cordon, and Johnson avoids this problem, but, as indicated above, at a cost.

When imported intermediate goods *are* produced at home, the foundations of the effective tariff concept seem to disappear

completely. If the strict unbiasedness and efficiency properties were the only reasons for using effective tariffs, the whole concept would have to be buried. I do not think, however, that this paper has discussed all of the properties of the effective tariff rate. In particular, it is suggested that the extent of bias in the effective tariff rate may be negligible. Thus, future research might reveal interesting properties that can be used to revitalize the effective tariff concept.

REFERENCES

1. B. BALASSA, "Tariff Protection in Industrial Countries: An Evaluation," *Jour. Pol. Econ.*, Dec. 1965, *73*, 573–94.
2. C. L. BARBER, "Canadian Tariff Policy," *Can. Jour. Econ. & Pol. Sci.*, Nov. 1955, *12*, 513–30.
3. G. BASEVI, "The U.S. Tariff Structure: Estimates of Effective Rates of Protection of U.S. Industries and Industrial Labor," *Rev. Econ. Stat.*, May 1966, *48*, 147–60.
4. W. M. CORDEN, "The Tariff," in A. Hunter (ed.), *The Economics of Australian Industry.* Parkville 1963.
5. ———, "The Structure of a Tariff System and the Effective Tariff Rate," *Jour. Pol. Econ.*, June 1966, *74*, 221–37.
6. O. A. DAVIS AND A. B. WHINSTON, "Welfare Economics and the Theory of Second Best," *Rev. Econ. Stud.*, Jan. 1965, *32*, 1–14.
7. H. JOHNSON, "The Theory of Tariff Structure with Special Reference to World Trade and Development," *Trade and Development.* Geneva 1965.
8. R. I. McKINNON, "Intermediate Products and Differential Tariffs: A Generalization of Lerner's Symmetry Theorem," *Quart. Jour. Econ.*, Nov. 1966, *80*, 584–615.
9. M. McMANUS, "Private and Social Costs in the Theory of Second Best," *Rev. Econ. Stud.*, July 1967, *34*, 317–22.

On Measuring the Nearness of Near-Moneys

By V. KARUPPAN CHETTY*

Monetary economists have long been concerned about the substitutability of the liabilities of various financial institutions for money. Knowledge of the degree of substitutability of such liquid assets for money is essential for many reasons. For example, if these assets are close substitutes for money, then the financial intermediaries can, in principle, reduce the effectiveness of any given monetary action. This is, in fact, the position taken by Gurley and Shaw [11] [12] [13] [14]. They argue that the monetary authorities did not succeed in reducing the liquidity in the economy during the postwar period due to the rapid growth of liabilities of financial intermediaries. There has been some theoretical discussion about the relevance of Gurley and Shaw's arguments, but as pointed out by Johnson [16] and Cagan [5], in a slightly different context, this issue cannot be resolved by theoretical arguments. In order to determine whether consumers regard the various liquid assets as substitutes for money or not, one has to look at their market behavior. The question is essentially an empirical one.

Recently, Feige [9], Hamburger [15], and Lee [20] attempted to test empirically the validity of Gurley and Shaw's hypothesis. Not surprisingly, these econometric studies are not in agreement about the substitutability of liquid assets for money. Feige [9], using temporal cross-section data of liquid asset holdings by states in the United States for the period 1949–59, found that the yields on nonbank intermediary liabilities did not affect the demand for money, defined as demand deposits plus currency. Hence he concluded that these assets are not substitutes for money. Lee [20], using Feige's and other types of data, concluded that savings and loan association shares are much better substitutes for money than time deposits. Hamburger [15], using U.S. time series data, found that the predictive power of the demand function for money did not increase significantly when the definition of money was expanded to include other liquid assets. For reasons given in the next section, the methods used by Feige [9] and Lee [20] are better than that of Hamburger [15].

In order to determine whether the public regards various liquid assets as substitutes for money, one has to determine empirically the shape of consumers' indifference curves for money and other liquid assets. In the present paper, a utility function which generates a variety of indifference curves is used to estimate directly the elasticity of substitution between money and other liquid assets. In the past, monetary economists attempted to test substitutability hypotheses by estimating the various (cross) interest elasticities of demand for money. In this paper, the various elasticities of substitution are used to test the same hypothesis. In theory, one would expect the two methods to lead to the same conclusion. However, the empirical results of our paper differ to a great extent from

* The author is associate professor of economics, Columbia University. He wishes to thank Milton Friedman, G. S. Becker, P. Cagan, A. G. Hart, K. Lancaster, R. Alcaly, and the referee for their comments and criticism. The paper also benefited from the discussion and comments from members of the Monetary Workshop of Columbia University. Mr. E. Wilcox provided valuable research assistance.

the findings of others and seem to be more reasonable in terms of our a priori expectations.

Furthermore, knowledge of these elasticities of substitution will be useful in answering many other questions in monetary economics. For instance, suppose one finds, as Lee [20] does, that some liquid assets are substitutes for money. In what ways can a monetary theorist, interested in controlling the liquidity of the economy, use these findings for policy purposes? If, for example, the supply of these liquid assets increases, other things being equal, their prices will fall. Using the demand function for money, the monetary theorist will determine the amount by which the demand for money will go down and recommend an appropriate policy measure for reducing the quantity of money. Thus policy action is taken only after the effect of the increases in other liquid assets shows up in the yields of these assets, which, of course, takes some time. On the other hand, using the elasticity of substitution approach one can immediately find the money equivalent of the change in other liquid assets and determine the amount by which the money supply should be reduced to maintain the same level of liquidity in the economy. This will avoid some delay in taking appropriate monetary measures which is certainly desirable, since many economists in the past have argued that there is considerable lag between the time a policy measure is taken and the time its effect is realized.

Another related problem, which has attracted the attention of monetary economists for many years, is the definition of money. The commonly used definition classifies demand deposits and currency as money. Friedman and Meiselman [10] and Cagan [5] define money as demand deposits, currency, and time deposits in commercial banks. One of the reasons for inclusion of time deposits is that they can not be meaningfully separated from demand deposits until the 1930's. Friedman and Meiselman also argue that "they are such close substitutes for other monetary items that it is preferable to treat them as if they were *perfect substitutes* than to omit them."

In reality, the various liquid assets may not be regarded as perfect substitutes for money nor can they be treated as completely unrelated to money. Hence the all-or-nothing approach in defining money does not seem to be very useful. Gurley [11], in this context, points out, "If the degree of substitutability between each type of monetary asset and money were known, liquid assets could be weighted in such a way that the constancy of this weighted amount would imply constant interest rates, other things the same." For illustrative purposes, Gurley used a definition which assigned weights of one to currency and demand deposits and weights of one-half to other liquid assets. It has been pointed out by several writers that the "best" set of weights can be derived using canonical correlation techniques or the method of principal components. This will give the "best" definition of money in some statistical sense, like maximum correlation or maximum variance, but it is difficult to give economic interpretation to these weights. Instead, in this paper, we suggest a method of aggregating the liquid assets, using the elasticities of substitution and other economically meaningful parameters.

Another related and controversial issue in monetary theory is concerned with the choice of the appropriate interest rate or rates to explain the demand for money. Some economists, such as Eisner [8] and Latane [19] maintain, following the Keynesian tradition of relating income and investment to long rates, that the relevant rate is that on long-term bonds. Bronfenbrenner and Mayer [4], Laidler [17]

and a few others advocate the use of short-term interest rates, since this reflects the opportunity cost of holding money. Gurley and Shaw [11] [12] [13] [14], naturally, argue for their candidate, the yield on nonbank intermediary liabilities. Lee [20] has recently presented empirical evidence for the superiority of the yield on saving and loan association shares over others but his conclusions are questionable for reasons pointed out in the sections that follow. Christ [6] and Lee [20] have used the relative quantities of these assets to form a weighted average of the various interest rates.

Regarding the choice of the interest rate, Turvey [22] remarked "Their relative importance depends upon the relative substitutability of long-term and short-term paper assets for money and upon their relative quantities. The former is unknown (hence all the argument), while the latter is measurable. . . . " Since the elasticities of substitution of various assets for money are estimated in the present paper, they are used to construct an index of interest rates as suggested by Turvey [22]. Since the demand for money in theory is a function of a number of highly correlated interest rates, it seems appropriate to use economically meaningful weights to come up with an average, rather than trying to choose *the* interest rate using some statistical criterion, like standard errors.

The plan of the paper is as follows. In Section I, the method of estimating the various elasticities of substitution, aggregation of the liquid assets, and construction of the interest rate index are discussed when assets are taken two at a time. Empirical estimates are presented for U.S. time series data for the period 1945–66. A more general method to handle all the assets simultaneously and the empirical results are presented in Section II. A new series based on the new definition of money and an index of interest rates are computed, and a velocity series based on the new definition of money is calculated and compared with other velocity series. Concluding remarks are made in Section III.

I. *A Model for Estimating the Elasticity of Substitution*

Throughout this paper, the term money (M) will be used to denote hand-to-hand currency plus demand deposits in commercial banks. To start with, it is assumed that there is one other financial asset, namely time deposits (T). Following the new approach to consumer demand theory, as developed by Becker [3] and Lancaster [18], let us assume that the consumer combines money and time deposits to produce various characteristics like liquidity, store of value, etc. We assume that the consumer combines M and T such that, for any given budget, he maximizes his satisfaction. The possibility of substituting T for M arises for two reasons: (1) M and T may have some common characteristics; (2) even if they have no characteristics in common, there may be substitution between different characteristics. The indifference curves between M and T may assume various possible shapes ranging from straight lines, in the case of perfect substitutes, to right angle curves, when they are consumed in fixed proportions. The degree of curvature of these indifference curves is a measure of substitutability of money and time deposits. To answer this question, we need a utility function, which generates a variety of indifference curves with different degrees of substitution and whose parameters can be estimated with the use of observed variables like quantities and prices. One such function which generates a variety of indifference curves, but has only few parameters, is the constant elasticity of substitution production function introduced by Arrow, *et al.* [2]. The CES function was introduced to study the degree of substitution between capital and

labor, but we can make use of this function to study the substitutability of money and other financial assets. The utility function of the consumer can be written as

$$U = (\beta_1 M^{-\rho} + \beta_2 T^{-\rho})^{-(1/\rho)}$$

where M and T represent money holdings and money value of time deposits in the next period respectively. Since this is a fairly general function, it can be assumed that it will provide an adequate approximation to the true utility function. Since only ordinal utility is used, β_1 can be assumed to be equal to 1 without loss of generality. This normalization is adopted throughout this paper. The implications of this normalization for the aggregation procedure are discussed later.

Suppose the consumer has cash holdings of M_0 dollars and wants to allocate them between M and T. If T represents the cash value of time deposits in the *next* period and if i is the rate of interest on time deposits of the current period, then the budget constraint of the consumer can be written as

$$M_0 = M + \frac{T}{1+i} \tag{1}$$

The slope of the budget line is $-(1+i)$. Hence $(1+i)$ can be considered as the ratio of the prices of money to time deposits.[1]

When the utility function is maximized subject to the budget constraint (1), the following marginal conditions are obtained:

$$\frac{\partial U}{\partial M} = \lambda \tag{2}$$

$$\frac{\partial U}{\partial T} = \lambda/(1+i) \tag{3}$$

$$M_0 = M + \frac{T}{(1+i)}, \tag{4}$$

where λ is the Lagrange multiplier. Dividing (2) by (3) we have

$$\frac{\beta_1}{\beta_2}\left(\frac{M}{T}\right)^{-\rho-1} = 1+i$$

Taking logarithms on both sides, rearranging terms, and adding a disturbance term, we have the regression model

$$log\frac{M}{T} = -\frac{1}{1+\rho}log\frac{\beta_2}{\beta_1} + \frac{1}{1+\rho}log\frac{1}{1+i} + e \tag{5}$$

Using data relating to M, T, and i, and making the usual assumptions about the disturbance term, we can estimate $1/1+\rho$ and the intercept term using least squares methods. From the intercept term, an estimate of β_2/β_1 can be obtained, and using a normalization rule β_1 and β_2 can then be estimated. The elasticity of substitution between money and time deposits is given by $\sigma = 1/1+\rho$. Thus whether any particular financial asset is a substitute or not can be directly tested using this regression.

This model has some similarities to the one used by Christ [6], since he regressed ratios of various stocks on interest rates and income, but his regression equations were not derived from any specific model like the one set out in this paper.

In our model,[2] the ratio of prices of

[1] Alternatively, one can assume that the consumer maximizes $U[T(1+i), M]$, subject to the constraint $M+T=M$, where T now represents the amount of time deposits in the *current* period. In both cases, the marginal conditions are the same. Also if one assumes that the consumer maximizes $U(Ti, M)$ subject to $M+T=M$, one obtains very similar results empirically, since the form of U is quite general.

[2] Milton Friedman, in a personal communication, pointed out to me that the price ratio has units of time in it and hence the estimated elasticity substitution will depend on the arbitrary choice of the time unit. To get around this difficulty, we should maximize the utility function subject to one more constraint on the flow of income from these assets. Where we have three or more assets, this can be done. If μ is the lagrangean multiplier

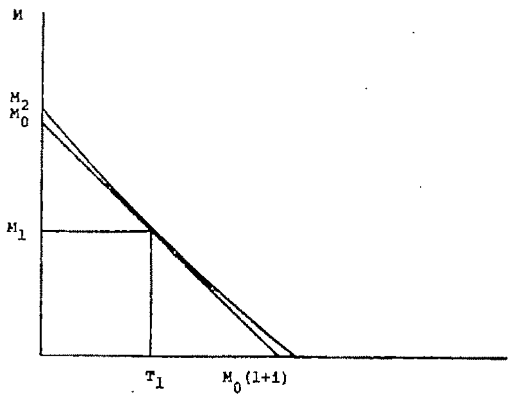


FIGURE 1. INDIFFERENCE CURVE BETWEEN M AND T

M and T is $1+i$. There are alternative ways of specifying the price of holding a dollar of M or T. If one takes the alternative cost approach, the price of holding a dollar of M will be the alternative income forgone. If, for example, the interest rate on corporate bonds, i_β, represents the alternative income, then the price ratio of T and M is given by $(i_\beta - i_T)/i_\beta = 1 - i_T/i_\beta$, where i_T is the interest rate on time deposits. If this price ratio is used in equation (5) this would imply that M/T will not change for a given percentage change in both i_T and i_β. This does not seem reasonable.

The empirically determined indifference curves can now be used to aggregate money and time deposits. The aggregation procedure is illustrated in Figure 1. For a given set of M_1 and T_1, we find the indifference curve that passes through that point. This is the maximum satisfaction that can be produced with this combination. From the same figure, we also find that the same satisfaction can be produced by using M_2 units of money alone. Hence adding T_1 units of time deposits to M_1 is equivalent to adding $(M_2 - M_1)$ units of money to M_1. Determination of such an M_2 is possible whenever the indifference curves intersect or are tangent to the M-axis. This measure of aggregate liquid assets *is exactly* the normalized utility function with $\beta_1 = 1$.

Friedman and Meiselman [10], in this context, point out that one should find out the moneyness of other liquid assets and use that to give an extended definition of money. Our aggregation procedure is along the same lines. In order to use this procedure, one need not determine completely the indifference curve that passes through a given M or T. Instead, if we assume that $\beta_1 = 1$, then the adjusted money, (M_a), is given by

$$M_a = (M^{-\rho} + \beta_2 T^{-\rho})^{-1/\rho}$$

after determining the degree of substitution between M and T. Our interest is to define the aggregate money. For this purpose, it seems quite natural to adopt the normalization rule, $\beta_1 = 1$.

If money and time deposits are identical commodities, then it follows that $\beta_2 = 1$ and $\rho = -1$. Then, $M_a = (M + T)$. Thus, the simple addition of M and T is justified, when M and T are identical. Of course, we do not have to make use of a CES function to reach this obvious conclusion. The fact, however, that the aggregation procedure reduces to a reasonable method under simple conditions increases our confidence in the procedure.

The methods discussed above were used to estimate the elasticity of substitution between money and commercial bank time deposits using time series data for the United States for the period 1945–66.[8]

for the new constraint, then the regressions will be of the form $\log M/T = a + b \log (1 + \mu i/\lambda)$. μi is now independent of the units of time. Here T is the amount of time deposits in the current period. Since μ/λ is the same for all regressions, b will be affected in the same manner in all regressions when i is small. Hence we can still run the same regression. Discussions with Gary Becker on this point have been very useful.

[8] The money stock and time deposits in commercial banks are from various issues of the *Federal Reserve Bulletin* (*FRB*). Interest rates on time deposits are from Cagan [5], for 1945–60 and from various issues of the *FRB* for 1961–66. Current income figures are from various issues of *Survey of Current Business*.

$$\log M/T = 1.510 + 34.69 \log 1/(1+i)$$
$$(1.569)$$

$$R^2 = .981$$

$$D.W. = .57$$

The elasticity of substitution between money and time deposits is significantly different from zero at the 5 percent level and considerably large. Hence we can treat M and T as very good substitutes for each other. Since there is evidence of autocorrelation among the disturbances, the standard error has been corrected using Wold's [24] method. When the logarithm of the current income is used as an additional explanatory variable, it turned out to be insignificant at the conventional levels. This is in agreement with our theoretical formulation, since our utility function implies that M/T will not depend on M_0 or income for given i. The implied estimate of β_2 is found to be

$$\beta_2 = \exp(-1.510/34.69) = .957.$$

The adjusted stock of money is given by

$$M_a = (M^{.971} + .957T^{.971})^{1.03}$$

An indifference curve between M and T is shown in Figure 1. This looks almost like a straight line, supporting Friedman's hypothesis that M and T are perfect substitutes.

The series relating to M, $M+T$, and M_a are shown in Table 1. It can be seen from columns (2) and (3) that there is not much difference between $M+T$ and M_a. Hence if we are interested in including T, we can simply add T to M.

Similar regressions were run to estimate the elasticities of substitution between money and the liabilities of savings and loan association (SL) and deposits in mutual savings banks (MS) for the period 1945–66.[4] The estimated equations are

TABLE 1—ADJUSTED MONEY STOCK BASED ON M AND T

M	$M+T$	M_a
102.3	132.4	133.5
110.0	143.8	145.0
113.6	148.8	150.1
111.6	147.4	148.6
111.2	147.3	148.5
117.7	154.0	155.3
124.5	162.0	163.8
129.0	169.7	171.2
130.5	174.2	175.8
134.4	181.2	182.9
138.2	186.6	188.4
139.7	190.3	192.4
138.6	194.7	196.9
144.2	207.4	209.8
145.6	212.2	214.8
144.7	216.8	219.7
149.4	212.2	214.4
151.6	248.3	252.1
157.3	268.3	272.6
164.0	289.2	294.0
172.0	317.2	322.8
175.8	335.4	341.9

$$\log M/SL = 4.612 + 101.851 \log 1/(1+i_{SL})$$
$$(8.900)$$

$$\bar{R}^2 = .946, \qquad D.W. = .510$$

$$\log M/MS = 2.297 + 27.637 \log 1/(1+i_{MS})$$
$$(1.354)$$

$$\bar{R}^2 = .980, \qquad D.W. = .829$$

i_{SL} and i_{MS} are yields on SL and MS respectively. The elasticity of substitution between M and SL is 101.851, the highest among the three assets. These findings strongly support Gurley and Shaw's [14] hypothesis and the findings of Lee [20], that savings and loan association shares are good substitutes for money. Hence we join Lee [20] in rejecting the conclusion of Feige [9] and Hamburger [15]. Feige, in fact, found that savings and loan shares are likely to be complementary to money.

Lee also found that, in the presence of the yield on nonbank intermediary liabilities, the yields on other assets, like time deposits, turn out to be statistically insignificant explanatory variables in the demand function for money. Hence this

[4] Savings and loan association shares and time deposits in mutual savings banks and their yields are from the *Savings and Loan Fact Book*, 1967.

implies that time deposits are not substitutes for money while savings and loan association shares are. This conclusion is somewhat questionable because the yields on various assets are highly correlated and hence the standard errors become unreliable. So one cannot conclude on the basis of "t" values which asset is a substitute and which is not.[5] Also Hamburger [15] reports that inclusion of other financial assets in M does not increase the R^2, but sometimes decreases it. But this is small wonder because the variance of $(M+T)$ is different from the variance of M. Since the dependent variables are different in the two regressions, R^2 can not be reliably used as a basis of comparison.

A weakness in our approach in estimating the degree of substitution is that we consider only one asset, in addition to M, at a time. The movements of M/T may reflect a shift between M and T or between one of them and other financial assets. Hence our estimates of the coefficients are likely to be biased. To remove this difficulty, we have to introduce all the assets simultaneously in our model and estimate the parameters. This is done in the next section. Briefly the results are: (1) The elasticity of substitution between T and M remains the same as before. (2) The substitutability of savings and loan association shares becomes somewhat weaker than before, but still they remain substitutes for money. (3) Deposits in mutual savings banks are also substitutes for money.

II. *Some Extensions of the Model*

We now assume that there are more than two financial assets. Let the utility function be

$$U = f(X_1, X_2, X_3 \cdots X_n),$$

where $X_1, X_2, \cdots, X_n$ are the various assets. We must now specify the form of the utility function. The purpose of our study will be defeated if we use the CES function for n inputs as generalized by Uzawa [23], since the partial elasticity of substitution between any two inputs is the same. The generalized CES function of Mukerjee [21] and Dhrymes and Kurz [7] does not assume that the partial elasticities are the same. Hence we assume that the utility function is given by (8). We assume that the budget constraint is given by (9) where Y is income and the r's are the yields on the various assets.

If the utility function is maximized subject to (9), we have the following first order conditions.

$$\frac{\partial U}{\partial M} - \lambda = 0 \tag{10}$$

$$\frac{\partial U}{\partial X_j} - \lambda/(1 + r_j) = 0 \tag{11}$$

$$(j = 1, 2, \cdots, n),$$

where λ is a Lagrangian multiplier.

The parameters of the utility function can be estimated using these conditions along the lines suggested by Dhrymes and Kurz [7]. Dividing (10) by (11) and manipulating as before, we have equation (12). If we substitute (12) in (9), we get an implicit relation between the interest rates, M, and Y. This relation can now be solved for M. Let the explicit relation between M, interest rates, and Y be given by

$$M = f(r_1, r_2, \cdots, r_n, Y).$$

Assuming log M has a valid Taylor series

[5] Lee has used the differentials between yields on liquid assets and the yield on money in his regressions to avoid the problem of collinearity. Michael J. Hamburger told me that when he recomputed the regressions after introducing the yield on money explicitly, instead of using the differentials, the interest elasticities for SL go down considerably.

$$U = (\beta M^{-\rho} + \beta_1 X_1^{-\rho_1} + \beta_2 X_2^{-\rho_2} \cdots + \beta_n X_n^{-\rho_n})^{-1/\rho} \tag{8}$$

$$M_0 = f(Y, r_1, r_2, \cdots, r_n) = M + \frac{X_1}{1 + r_1} + \frac{X_2}{1 + r_2} + \frac{X_n}{1 + r_n} \tag{9}$$

$$\log X_j = \frac{-1}{\rho_j + 1} \log \frac{\beta\rho}{\beta_j\rho_j} - \frac{1}{\rho_j + 1} \log \frac{1}{1+r_j} + \frac{\rho + 1}{\rho_j + 1} \log M, \tag{12}$$

$$\log M = a_0 + \sum_{j=1}^{n} a_j \log (1 + r_j) + a_{n+1} \log Y. \tag{13}$$

---

expansion in terms of log r_j and log Y, we can write this expansion as equation (13). When the disturbance terms are introduced in equations (12) and (13), the parameters of the system can be estimated using two-stage least squares. First, log M is estimated as a function of Y and the r_j's, which are assumed to be exogenous. The estimate of log M is then inserted in (12) and each equation is estimated individually. The estimates obtained in the second stage are consistent and can be used to determine the implied estimates of the parameters of the production function. Also the asymptotic variances of these estimates can be calculated using standard methods of large sample distribution theory.[6] These estimates can then be used to calculate the partial elasticity of substitution between M and other assets defined as:

$$\sigma_{M,X_j} = \frac{d \log (M/X_j)}{d \log \left(\frac{\partial U/\partial M}{\partial U/\partial X_j}\right)} = \frac{1}{(1+\rho)+(\rho_j-\rho) \Big/ \left[1 + \frac{\beta_j\rho_j X_j^{-\rho_j}}{\beta\rho M^{-\rho}}\right]} \tag{14}$$

This is the Hicks-Allen [1] direct partial elasticity of substitution.

Here income and interest rates may not be strictly exogenous variables[7] and hence the two-stage least squares may not be consistent for equation (12). Thus the ordinary least squares may be no worse than the two-stage least squares as far as the inconsistency is concerned. Using the least squares method, one can simply estimate (12) and omit equation (13) and the Taylor series approximation altogether. Both the methods were tried in estimating the regression. Since the estimates of the parameters were almost identical, only the ordinary least squares estimates are given below.

The estimated regression equations for time deposits, savings and loan associa-

[6] Dhrymes and Kurz [7] suggest a method of constructing the confidence interval for ρ_i, given the confidence interval for $1/(\rho_i-1)$. This is incorrect because if $1/(\rho_i-1)$ has a "t" distribution, the moments of ρ_i will not in general exist for finite samples. Hence only asymptotic variance can be used. Also in this case, the distribution of $1/(\rho_i-1)$, a two-stage least squares estimate, is not known for finite samples.

[7] This was pointed out to me by a referee.

$$
(15)\quad \begin{cases} \log T = .118 - \underset{(2.69)}{40.384} \log P_T + \underset{(.19)}{.649} \log M & \bar{R}^2 = .991, \quad D.W. = .87 \\[2ex] \log SL = -15.398 - \underset{(10.25)}{52.346} \log P_{SL} + \underset{(.49)}{3.517} \log M & \bar{R}^2 = .983, \quad D.W. = .49 \\[2ex] \log MS = -3.517 - \underset{(3.32)}{24.105} \log P_{MS} + \underset{(.197)}{1.217} \log M & \bar{R}^2 = .992, \quad D.W. = .84 \end{cases}
$$

$$
(16)\qquad M_a = [M^{.954} + 1.020T^{.975} + .880MS^{.959} + .616SL^{.981}]^{1.026}
$$

---

tion shares, and mutual savings bank deposits are given in equation (15), where $P_j = 1/(1+r_j)$.

The partial elasticities of substitution between M and other assets at their mean values are calculated using (14). We have the following results:

$$\sigma_{M,T} = 30.864$$
$$\sigma_{M,SL} = 35.461$$
$$\sigma_{M,MS} = 23.310$$

The partial elasticities of substitution, σ_{MT} and $\sigma_{M,MS}$ are similar to those obtained in the previous section. The partial elasticity of substitution between M and SL is 35.461, compared to 101.851 obtained before. Our second estimate of $\sigma_{M,SL} = 35.461$ seems more reasonable than 101.851, since the former is close to $\sigma_{M,T}$ and $\sigma_{M,MS}$. Also it is free from the bias due to the single equation approach mentioned earlier. There is no reason why $\sigma_{M,SL}$ alone should be considerably different from those of other financial assets. Also in view of the larger standard errors associated with the regression equation relating to SL, one can not conclude that the liabilities of savings and loan association are the best substitutes for money. The conclusion that can be drawn from these regressions is that T, SL, and MS are all good substitutes for money. Thus, we have provided empirical evidence for Gurley and Shaw's hypothesis.

Also the utility function in (8) can be used to determine the adjusted money stock as before. The adjusted stock of money in this case is given by equation (16). It is interesting to observe that all the exponents are very close to one. Also the coefficient of time deposits is approximately 1, which again supports Friedman's definition. Approximately we can write M as

$$M_a' = M + T + .880MS + .615SL$$

On the basis of the coefficients of various assets we can probably rank the assets in the decreasing order of closeness to money as follows: T, MS, and SL. Thus our conclusions are different from those of Lee [20].

It is comforting to find that the weights for T, MS, and SL are in the range (0, 1). Since the estimates are derived from a least squares regression with the unrestricted coefficients, the results are really amazing. We believe it is only fair to remark that it is not a very common event in econometrics to find the estimated coefficients within the relevant range. The weight for T is slightly greater than 1; but in view of the standard error of the coefficients of the regression relating to T and M, this is not bad. Also, if T is almost as good as

TABLE 2—ADJUSTED MONEY STOCK BASED ON M, T, SL AND MS.

	M_a	M'_a	$V_1=Y/M$	$V_2=$ $Y/M+T$	$V_3=Y/Ma$	i (percent)
1945	144.75	150.42	2.09	1.61	1.47	1.56
1946	157.67	163.81	1.92	1.47	1.33	1.48
1947	164.16	170.49	2.06	1.57	1.42	1.48
1948	164.38	170.36	2.32	1.76	1.57	1.53
1949	160.20	171.97	2.32	1.75	1.55	1.59
1950	173.91	180.21	2.42	1.85	1.63	1.66
1951	183.96	190.69	2.64	2.03	1.78	1.74
1952	194.69	201.40	2.69	2.04	1.78	1.94
1953	203.33	209.69	2.80	2.10	1.79	2.03
1954	215.01	221.22	2.70	2.00	1.68	2.13
1955	225.02	231.16	2.88	2.13	1.68	2.19
1956	233.78	239.52	3.00	2.20	1.78	2.36
1957	243.53	248.36	3.18	2.26	1.80	2.67
1958	262.37	266.84	3.07	2.13	1.68	2.76
1959	272.71	276.58	3.32	2.27	1.76	2.96
1960	284.23	286.94	3.47	2.31	1.76	3.20
1961	306.57	308.51	3.47	2.24	1.68	3.32
1962	333.52	333.97	3.76	2.26	1.67	3.62
1963	364.39	363.70	3.73	2.28	1.60	3.76
1964	396.45	394.81	3.86	2.19	1.58	3.85
1965	433.73	431.15	3.98	2.16	1.56	4.00
1966	457.43	453.85	4.23	2.22	1.60	4.34
Mean velocity			$\bar{V}_1=2.99$	$\bar{V}_2=2.03$	$\bar{V}_3=1.65$	
Coefficient of Variation			21.77	12.1	7.6	

money, as Friedman and others suspect, we would even expect the weight to be slightly greater than 1, since it yields a positive rate of return.

Since we have estimates of substitution and other parameters, we can define an index of interest rates. The exponents in the equation for M_a are very close to 1, and we will just use the marginal rates of substitution between M and other assets, namely 1, .88, and .615. Hence the interest rate index is given by

$$i_a = (i_T + .88 i_s + .615 i_{SL}) \div (1 + .88 + .615)$$

The adjusted money stock series M_a and M'_a, the interest index, i_a, and velocities based on various definitions of money are given in Table 2. It can be seen that there is not much difference between M_a and M'_a. The correlation coefficient beween M_a and $M+T$ is .999. This does not mean that one definition is as good as the other for all purposes. The mean and standard deviation of $M+T$ are 202.67 and 57.73 respectively, while the mean and standard deviation of M_a, are respectively, 253.72 and 93.07. There is considerable difference between the standard deviations. Hence for purposes of controlling or explaining money supply or demand, the two series will have different implications. But if one is using money stock to predict some other variable, say national income, then $M+T$ is as good as M'_a, or for that matter as M.

Also it can be seen from Table 2 that velocity based on M'_a is virtually a constant, while the other velocities vary to a considerable extent. The coefficient of variation is about one-third of that based on the traditional definition of money and about one-half of that based on Friedman's definition of money. Thus, as Gurley and Shaw and others have argued,

we can attribute the postwar rise in velocity (based on the traditional definition of money) to the increased availability of money substitutes.

III. *Summary*

In this paper, some methods of estimating the degree of substitution between money and other financial assets are set out and used to test the substitution hypothesis of Gurley and Shaw using U.S. time series data for the period 1945–66. Also a method of aggregating the various liquid assets and interest rates using the elasticities of substitution and other economically meaningful parameters is presented. The empirical findings of this paper are: (1) Time deposits and savings and loan association shares are also very good substitutes for money and can be ranked in decreasing or order of closeness to money as follows: T, MS, and SL. (2) Our empirical results lead to the following definition of money:

$$M'_a = M + T + .880MS + .615SL$$

(3) Velocity, based on the new definition of money, has been virtually a constant since 1951.

References

1. R. G. D. Allen and J. R. Hicks, "A Reconsideration of the Theory of Value, II," *Economica*, May 1934, *1*, 196–219.
2. K. J. Arrow, H. B. Chenery, B. S. Minhas, and R. M. Solow, "Capital-Labor Substitution and Economic Efficiency", *Rev. Econ. Stat.*, Aug. 1961, *63*, 225–50.
3. G. S. Becker, "A Theory of the Allocation of Time", *Econ. Jour.*, Sept. 1965, *75*, 493–517.
4. M. Bronfenbrenner and T. Mayer, "Liquidity Functions in the American Economy," *Econometrica*, Oct. 1960, *28*, 810–34.
5. P. Cagan, *Determinants and Effects of Changes in the Stock of Money, 1875–1960*, New York, 1965.
6. C. F. Christ, "Interest Rates and Portfolio Selection among Liquid Assets in the U.S.", *Measurement in Economics: Studies in Mathematical Economics and Econometrics in Memory of Yehuda Greenfeld*, Stanford 1963.
7. P. J. Dhrymes and M. Kurz, "Technology and Scale in Electricity Generation," *Econometrica*, July 1964, *32*, 287–315.
8. R. Eisner, "Another Look at Liquidity Preference", *Econometrica*, July 1963, *31*, 531–38.
9. E. L. Feige, *The Demand for Liquid Assets: A Temporal Cross-Section Analysis*. Englewood Cliffs 1964.
10. M. Friedman and D. Meiselman, "The Relative Stability of Monetary Velocity and the Investment Multiplier in the United States, 1897–1958," in Commission on Money and Credit, *Stabilization Policies*, Englewood Cliffs 1963.
11. J. G. Gurley, "Liquidity and Financial Institutions in the Postwar Economy," U.S. Congress, Joint Economic Committee, *Study of Employment, Growth and Price Levels*, Study paper No. 14, Jan. 1960.
12. J. G. Gurley and E. S. Shaw, "Financial Aspects of Economic Development," *Am. Econ. Rev.*, Sept. 1955, *45*, 515–38.
13. ———, "Financial Intermediaries and the Saving-Investment Process," *Jour. Fin.*, May 1956, *11*, 257–76.
14. ———, *Money in a Theory of Finance*. Washington, D.C. 1960.
15. M. J. Hamburger, "The Demand for Money by Households, Money Substitutes, and Monetary Policy," *Jour. Pol. Econ.*, Dec. 1966, *74*, 600–23.
16. H. G. Johnson, "Monetary Theory and Policy," *Am. Econ. Rev.*, June 1962, *52*, 335–84.
17. D. Laidler, "Some Evidence on the Demand for Money—Some Empirical Evidence," *Jour. Pol. Econ.*, Dec. 1966, *74*, 543–55.

18. K. J. LANCASTER, "A New Approach to Consumer Theory," *Jour. Pol. Econ.*, April 1966, *74*, 132–57.
19. H. A. LATANE, "Cash Balances and the Interest Rate—A Pragmatic Approach," *Rev. Econ. Stat.*, Nov. 1954, *36*, 456–60.
20. T. H. LEE, "Alternative Interest Rates and the Demand for Money: The Empirical Evidence," *Am. Econ. Rev.*, Dec. 1967, *57*, 1168–81.
21. V. MUKERJEE, "A Generalized S.M.A.C. Function with Constant Ratios of Elasticity of Substitution," *Rev. Econ. Stud.*, Oct. 1963, *30*, 233–36.
22. R. TURVEY, "On the Demand for Money," *Econometrica*, April 1965, *33*, 459–60.
23. H. UZAWA, "Production Functions with Constant Elasticities of Substitution," *Rev. Econ. Stud.*, Oct. 1962, 291–99.
24. H. WOLD, *Demand Analysis*. New York 1953.

Production, Consumption, and Externalities

By ROBERT U. AYRES AND ALLEN V. KNEESE*

> "For all that, welfare economics can no more reach conclusions applicable to the real world without some knowledge of the real world than can positive economics" [21].

Despite tremendous public and governmental concern with problems such as environmental pollution, there has been a tendency in the economics literature to view externalities as exceptional cases. They may distort the allocation of resources but can be dealt with adequately through simple *ad hoc* arrangements. To quote Pigou:

> When it was urged above, that in certain industries a wrong amount of resources is being invested because the value of the marginal social net product there differs from the value of the marginal private net product, it was tacitly assumed that in the main body of industries these two values are equal [22][1].

And Scitovsky, after having described his cases two and four which deal with technological externalities affecting consumers and producers respectively, says:

> The second case seems exceptional, because most instances of it can be and usually are eliminated by zoning ordnances and industrial regulations concerned with public health and safety. The fourth case seems unimportant, simply because examples of it seem to be few and exceptional [25].

We believe that at least one class of externalities—those associated with the disposal of residuals resulting from the consumption and production process—must be viewed quite differently.[2] They are a normal, indeed, inevitable part of these processes. Their economic significance tends to increase as economic development proceeds, and the ability of the ambient environment to receive and assimilate them is an important natural resource of increasing value.[3] We will argue below that

* The authors are respectively visiting scholar and director, Quality of the Environment Program, Resources for the Future, Inc. We are indebted to our colleagues Blair Bower, Orris Herfindahl, Charles Howe, John Krutilla, and Robert Steinberg for comments on an earlier draft. We have also benefited from comments by James Buchanan, Paul Davidson, Robert Dorfman, Otto Eckstein, Myrick Freeman, Mason Gaffney, Lester Lave, Herbert Mohring, and Gordon Tullock.

[1] Even Baumol who saw externalities as a rather pervasive feature of the economy tends to discuss external diseconomies like "smoke nuisance" entirely in terms of particular examples [3]. A perspective more like that of the present paper is found in Kapp [16].

[2] We by no means wish to imply that this is the only important class of externalities associated with production and consumption. Also, we do not wish to imply that there has been a lack of theoretical attention to the externalities problem. In fact, the past few years have seen the publication of several excellent articles which have gone far toward systematizing definitions and illuminating certain policy issues. Of special note are Coase [9], Davis and Whinston [12], Buchanan and Stubblebine [6], and Turvey [27]. However, all these contributions deal with externality as a comparatively minor aberration from Pareto optimality in competitive markets and focus upon externalities between two parties. Mishan, after a careful review of the literature, has commented on this as follows: "The form in which external effects have been presented in the literature is that of partial equilibrium analysis; a situation in which a single industry produces an equilibrium output, usually under conditions of perfect competition, some form of intervention being required in order to induce the industry to produce an "ideal" or "optimal" output. If the point is not made explicitly, it is tacitly understood that unless the rest of the economy remains organized in conformity with optimum conditions, one runs smack into Second Best problems" [21].

[3] That external diseconomies are integrally related to economic development and increasing congestion has been noted in passing in the literature. Mishan has commented: "The attention given to external effects in

the common failure to recognize these facts may result from viewing the production and consumption processes in a manner that is somewhat at variance with the fundamental law of conservation of mass.

Modern welfare economics concludes that if (1) preference orderings of consumers and production functions of producers are independent and their shapes appropriately constrained, (2) consumers maximize utility subject to given income and price parameters, and (3) producers maximize profits subject to the price parameters; a set of prices exists such that no individual can be made better off without making some other individual worse off. For a given distribution of income this is an efficient state. Given certain further assumptions concerning the structure of markets, this "Pareto optimum" can be achieved via a pricing mechanism and voluntary decentralized exchange.

If waste assimilative capacity of the environment is scarce, the decentralized voluntary exchange process cannot be free of uncompensated technological external diseconomies unless (1) all inputs are fully converted into outputs, with no unwanted material residuals along the way,[4] and all final outputs are utterly destroyed in the process of consumption, or (2) property rights are so arranged that all relevant environmental attributes are in private ownership and these rights are exchanged in competitive markets. Neither of these conditions can be expected to hold in an actual economy and they do not.

Nature does not permit the destruction of matter except by annihilation with antimatter, and the means of disposal of unwanted residuals which maximizes the internal return of decentralized decision units is by discharge to the environment, principally, watercourses and the atmosphere. Water and air are traditionally examples of free goods in economics. But in reality, in developed economies they are common property resources of great and increasing value presenting society with important and difficult allocation problems which exchange in private markets cannot resolve. These problems loom larger as increased population and industrial production put more pressure on the environment's ability to dilute and chemically degrade waste products. Only the crudest estimates of present external costs associated with residuals discharge exist but it would not be surprising if these costs were in the tens of billions of dollars annually.[5] Moreover, as we shall emphasize again, technological means for processing or purifying one or another type of waste discharge do not destroy the residuals but only alter their form. Thus, given the level, patterns, and technology of production and consumption, recycle of materials into productive uses or discharge into an alternative medium are the only general options for protecting a particular environmental medium such as water. Residual problems must be seen in a broad regional or economy-wide context rather

the recent literature is, I think, fully justified by the unfortunate, albeit inescapable, fact that as societies grow in material wealth, the incidence of these effects grows rapidly . . . " [21]; and Buchanan and Tullock have stated that as economic development proceeds, "congestion" tends to replace "co-operation" as the underlying motive force behind collective action, i.e., controlling external diseconomies tends to become more important than cooperation to realize external economies [7].

[4] Or any residuals which occur must be stored on the producer's premises.

[5] It is interesting to compare this with estimates of the cost of another well known misallocation of resources that has occupied a central place in economic theory and research. In 1954, Harberger published an estimate of the welfare cost of monopoly which indicated that it amounted to about .07 percent of GNP [15]. In a later study, Schwartzman calculated the allocative cost at only .01 percent of GNP [24]. Leibenstein generalized studies such as these to the statement that " . . . in a great many instances the amount to be gained by increasing allocative efficiency is trivial . . . " [19]. But Leibenstein did not consider the allocative costs associated with environmental pollution.

than as separate and isolated problems of disposal of gas, liquid, and solid wastes.

Frank Knight perhaps provides a key to why these elementary facts have played so small a role in economic theorizing and empirical research.

> The next heading to be mentioned ties up with the question of dimensions from another angle, and relates to the second main error mentioned earlier as connected with taking food and eating as the type of economic activity. The basic economic magnitude (value or utility) is service, not good. It is inherently a stream or flow in time . . . [18].[6]

Almost all of standard economic theory is in reality concerned with services. Material objects are merely the vehicles which carry some of these services, and they are exchanged because of consumer preferences for the services associated with their use or because they can help to add value in the manufacturing process. Yet we persist in referring to the "final consumption" of goods as though material objects such as fuels, materials, and finished goods somehow disappeared into the void—a practice which was comparatively harmless so long as air and water were almost literally free goods.[7] Of course, residuals from both the production and consumption processes remain and they usually render disservices (like killing fish, increasing the difficulty of water treatment, reducing public health, soiling and deteriorating buildings, etc.) rather than services. Control efforts are aimed at eliminating or reducing those disservices which flow to consumers and producers whether they want them or not and which, except in unusual cases, they cannot control by engaging in individual exchanges.[8]

I. *The Flow of Materials*

To elaborate on these points, we find it useful initially to view environmental pollution and its control as a materials balance problem for the entire economy.[9] The inputs to the system are fuels, foods, and raw materials which are partly converted into final goods and partly become waste residuals. Except for increases in inventory, final goods also ultimately enter the waste stream. Thus goods which are "consumed" really only render certain services. Their material substance remains in existence and must either be reused or discharged to the ambient environment.

In an economy which is closed (no imports or exports) and where there is no net accumulation of stocks (plant, equipment, inventories, consumer durables, or residential buildings), the amount of residuals inserted into the natural environment must be approximately equal to the weight of basic fuels, food, and raw materials entering the processing and production system, plus oxygen taken from the atmosphere.[10] This result, while obvious

[6] The point was also clearly made by Fisher: "The only true method, in our view, is to regard uniformly as income the *service* of a dwelling to its owner (shelter or money rental), the *service* of a piano (music), and the *service* of food (nourishment) . . . " (emphasis in original) [14].

[7] We are tempted to suggest that the word consumption be dropped entirely from the economist's vocabulary as being basically deceptive. It is difficult to think of a suitable substitute, however. At least, the word consumption should not be used in connection with goods, but only with regard to services or flows of "utility."

[8] There is a substantial literature dealing with the question of under what conditions individual exchanges can optimally control technological external diseconomies. A discussion of this literature, as it relates to waterborne residuals, is found in Kneese and Bower [17].

[9] As far as we know, the idea of applying materials balance concepts to waste disposal problems was first expressed by Smith [26]. We also benefitted from an unpublished paper by Joseph Headley in which a pollution "matrix" is suggested. We have also found references by Boulding to a "spaceship economy" suggestive [4]. One of the authors has previously used a similar approach in ecological studies of nutrient interchange among plants and animals; see [1].

[10] To simplify our language, we will not repeat this essential qualification at each opportunity, but assume it applies throughout the following discussion. In addition, we must include residuals such as *NO* and NO_2 arising from reactions between components of the air itself but occurring as combustion by-products.

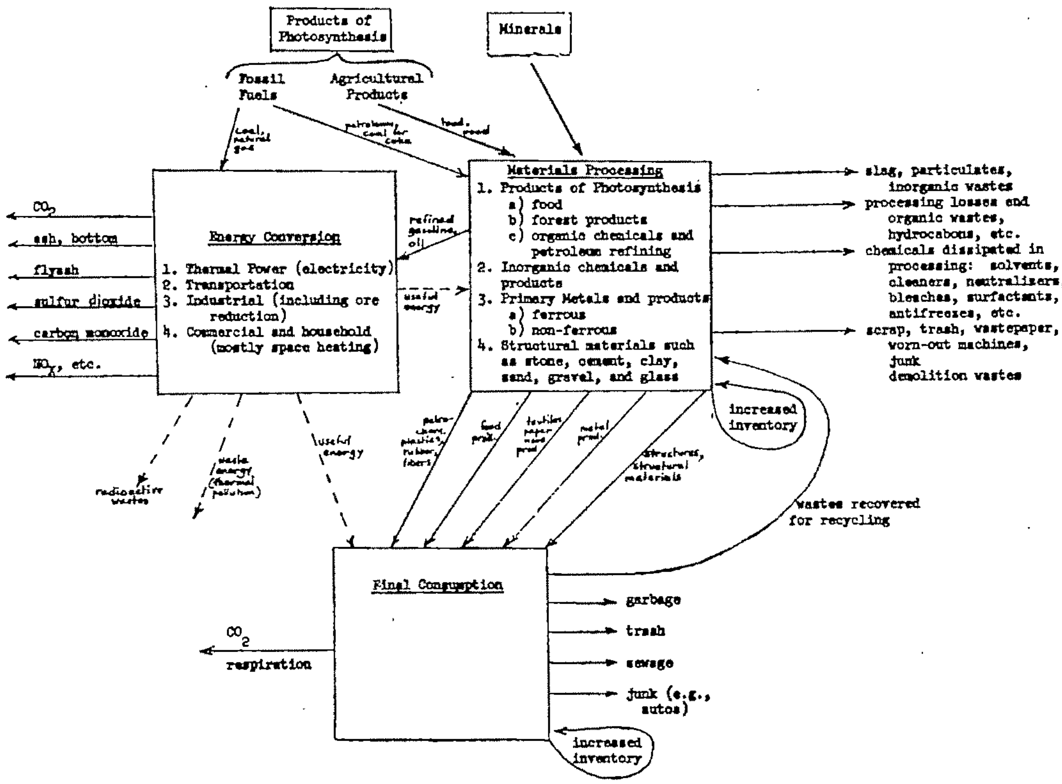


Figure 1.—Materials Flow

upon reflection, leads to the, at first rather surprising, corollary that residuals disposal involves a greater tonnage of materials than basic materials processing, although many of the residuals, being gaseous, require no physical "handling."

Figure 1 shows a materials flow of the type we have in mind in greater detail and relates it to a broad classification of economic sectors for convenience in our later discussion, and for general consistency with the Standard Industrial Classification. In an open (regional or national) economy, it would be necessary to add flows representing imports and exports. In an economy undergoing stock or capital accumulation, the production of residuals in any given year would be less by that amount than the basic inputs. In the entire U.S. economy, accumulation accounts for about 10–15 percent of basic annual inputs, mostly in the form of construction materials, and there is some net importation of raw and partially processed materials amounting to 4 or 5 percent of domestic production. Table 1 shows estimates of the weight of raw materials produced in the United States in several recent years, plus net imports of raw and partially processed materials.

Of the active inputs,[11] perhaps three-quarters of the overall weight is eventually discharged to the atmosphere as carbon (combined with atmospheric oxygen in the form of CO or CO_2) and hydrogen (combined with atmospheric oxygen as H_2O) under current conditions. This results from combustion of fossil fuels and from animal respiration. Discharge of carbon dioxide can be considered harmless in the short run. There are large "sinks" (in the form of vegetation and large water bodies,

[11] See footnote to Table 1.

TABLE 1—WEIGHT OF BASIC MATERIALS PRODUCTION IN THE UNITED STATES PLUS NET IMPORTS, 1963 (10^6 tons)

	1963	1964	1965
Agricultural (incl. fishery and wildlife and forest) products			
Food: Crops (excl. livestock feed)	125	128	130
Food: Livestock	100	103	102
Other products	5	6	6
Fishery	3	3	3
Forestry products (85 per cent dry wt. basis)			
Sawlogs	53	55	56
Pulpwood	107	116	120
Other	41	41	42
Total	434	452	459
Mineral fuels	1,337	1,399	1,448
Other minerals			
Iron ore	204	237	245
Other metal ores	161	171	191
Other nonmetals	125	133	149
Total	490	541	585
Grand total[a]	2,261	2,392	2,492

[a] Excluding construction materials, stone, sand, gravel, and other minerals used for structural purposes, ballast, fillers, insulation, etc. Gangue and mine tailings are also excluded from this total. These materials account for enormous tonnages but undergo essentially no chemical change. Hence, their use is more or less tantamount to physically moving them from one location to another. If this were to be included, there is no logical reason to exclude material shifted in highway cut and fill operations, harbor dredging, land-fill, plowing, and even silt moved by rivers. Since a line must be drawn somewhere, we chose to draw it as indicated above.

Source: R. U. Ayres and A. V. Kneese [2, p. 630].

mainly the oceans) which reabsorb this gas, although there is evidence of net accumulation of CO_2 in the atmosphere. Some experts believe that the latter is likely to show a large relative increase, as much as 50 per cent by the end of the century, possibly giving rise to significant —and probably, on balance, adverse— weather changes.[12] Thus continued combustion of fossil fuels at a high rate could produce externalities affecting the entire world. The effects associated with most residuals will normally be more confined, however, usually limited to regional air and water sheds.

The remaining residuals are either gases (like carbon monoxide, nitrogen dioxide, and sulfur dioxide—all potentially harmful even in the short run), dry solids (like rubbish and scrap), or wet solids (like garbage, sewage, and industrial wastes suspended or dissolved in water). In a sense, the dry solids are an irreducible, limiting form of waste. By the application of appropriate equipment and energy, most undesirable substances can, in principle, be removed from water and air streams[13]—but what is left must be disposed of in solid form, transformed, or reused. Looking at the matter in this way clearly reveals a primary interdependence between the various waste streams which casts into doubt the traditional classification of air, water, and land pollution as individual categories for purposes of planning and control policy.

Residuals do not necessarily have to be discharged to the environment. In many instances, it is possible to recycle them back into the productive system. The materials balance view underlines the fact that the throughput of new materials necessary to maintain a given level of production and consumption decreases as the technical efficiency of energy conversion and materials utilization increases. Similarly, other things being equal, the longer that cars, buildings, machinery, and other durables remain in service, the fewer new materials are required to compensate for loss, wear, and obsolescence—although the use of old or worn machinery (e.g., automobiles) tends to increase other residuals problems. Technically efficient combustion of (desulfurized) fossil fuels

[12] See [30]. There is strong evidence that discharge of residuals has already affected the climate of individual cities; see Lowry [20].

[13] Except CO_2, which may be harmful in the long run, as noted.

would leave only water, ash, and carbon dioxide as residuals, while nuclear energy conversion need leave only negligible quantities of material residuals (although thermal pollution and radiation hazards cannot be dismissed by any means).

Given the population, industrial production, and transport service in an economy (a regional rather than a national economy would normally be the relevant unit), it is possible to visualize combinations of social policy which could lead to quite different relative burdens placed on the various residuals-receiving environmental media; or, given the possibilities for recycle and less residual-generating production processes, the overall burden to be placed upon the environment as a whole. To take one extreme, a region which went in heavily for electric space heating and wet scrubbing of stack gases (from steam plants and industries), which ground up its garbage and delivered it to the sewers and then discharged the raw sewage to watercourses, would protect its air resources to an exceptional degree. But this would come at the sacrifice of placing a heavy residuals load upon water resources. On the other hand, a region which treated municipal and industrial waste water streams to a high level and relied heavily on the incineration of sludges and solid wastes would protect its water and land resources at the expense of discharging waste residuals predominantly to the air. Finally, a region which practiced high level recovery and recycle of waste materials and fostered low residual production processes to a far reaching extent in each of the economic sectors might discharge very little residual waste to any of the environmental media.

Further complexities are added by the fact that sometimes it is possible to modify an environmental medium through investment in control facilities so as to improve its assimilative capacity. The clearest, but far from only, example is with respect to watercourses where reservoir storage can be used to augment low river flows that ordinarily are associated with critical pollution (high external cost situations).[14] Thus internalization of external costs associated with particular discharges, by means of taxes or other restrictions, even if done perfectly, cannot guarantee Pareto optimality. Investments involving public good aspects must enter into an optimal solution.[15]

To recapitulate our main points briefly: (1) Technological external diseconomies are not freakish anomalies in the processes of production and consumption but an inherent and normal part of them. (2) These external diseconomies are quantitatively negligible in a low-population or economically undeveloped setting, but they become progressively (nonlinearly) more important as the population rises and the level of output increases (i.e., as the natural reservoirs of dilution and assimilative capacity become exhausted).[16] (3) They cannot be properly dealt with by considering environmental media such as air and water in isolation. (4) Isolated and *ad hoc* taxes and other restrictions are not sufficient for their optimum control, although they are essential elements in a more systematic and coherent program of environmental quality management. (5) Public investment programs, particularly including transportation systems, sewage disposal, and river flow regulation, are intimately related to the amounts and

[14] Careful empirical work has shown that this technique can fit efficiently into water quality management systems. See Davis [11].

[15] A discussion of the theory of such public investments with respect to water quality management is found in Boyd [5].

[16] Externalities associated with residuals discharge may appear only at certain threshold values which are relevant only at some stage of economic development and industrial and population concentrations. This may account for their general treatment as "exceptional" cases in the economics literature. These threshold values truly were exceptional cases for less developed economies.

effects of residuals and must be planned in light of them.

It is important to develop not only improved measures of the external costs resulting from differing concentrations and duration of residuals in the environment but more systematic methods for forecasting emissions of external-cost-producing residuals, technical and economic trade-offs between them, and the effects of recycle on environmental quality.

In the hope of contributing to this effort and of revealing more clearly the types of information which would be needed to implement such a program, we set forth a more formal model of the materials balance approach in the following sections and relate it to some conventional economic models of production and consumption. The main objective is to make some progress toward defining a system in which flows of services and materials are simultaneously accounted for and related to welfare.

II. *Basic Model*

The take off point for our discussion is the Walras-Cassel general equilibrium model,[17] extended to include intermediate consumption, which involve the following quantities:

$$\overbrace{r_1, \cdots\cdots\cdots, r_M}^{\text{resources and services}}$$

$$\overbrace{X_1, \cdots\cdots\cdots, X_N}^{\text{products or commodities}}$$

$$\overbrace{v_1, \cdots\cdots\cdots, v_M}^{\text{resource prices}}$$

$$\overbrace{p_1, \cdots\cdots\cdots, p_N}^{\text{product or commodity prices}}$$

$$\overbrace{Y_1, \cdots\cdots\cdots, Y_N}^{\text{final demands}}$$

[17] The original references are Walras [28] and Cassel [8]. Our own treatment is largely based on Dorfman *et al.* [13].

The M basic resources are allocated among the N sectors as follows:

$$\begin{aligned} r_1 &= a_{11}X_1 + a_{12}X_2 + \cdots + a_{1N}X_N \\ r_2 &= a_{21}X_1 + a_{22}X_2 + \cdots + a_{2N}X_N \\ &\vdots \\ r_M &= a_{M1}X_1 + a_{M2}X_2 + \cdots + a_{MN}X_N \end{aligned}$$

(1a) or

$$r_j = \sum_{k=1}^{N} a_{jk}X_k \qquad j = 1, \cdots, M$$

In (1a) we have implicitly assumed that there is no possibility of factor or process substitution and no joint production. These conditions will be discussed later. In matrix notation we can write:

$$[r_{j1}]_{M,1} = [a_{jk}]_{M,N} \cdot [X_{k1}]_{N,1} \tag{1b}$$

where $[a]$ is an $M \times N$ matrix.

A similar set of equations describes the relations between commodity production and final demand:

$$X_k = \sum_{l=1}^{N} A_{kl}Y_l \qquad k = 1, \cdots, N \tag{2a}$$

$$[X_{k1}]_{N,1} = [A_{kl}]_{N,N} \cdot [Y_{l1}]_{N,1} \tag{2b}$$

and the matrix $[A]$ is given by

$$[A] = [I - C]^{-1} \tag{3}$$

where $[I]$ is the unit diagonal matrix and the elements C_{ij} of the matrix $[C]$ are essentially the well known Leontief input coefficients. In principle these are functions of the existing technology and, therefore, are fixed for any given situation.

By combining (1) and (2), we obtain a set of equations relating resource inputs directly to final demand, viz.,

$$\begin{aligned} r_j &= \sum_{k=1}^{N} a_{jk} \sum_{l=1}^{N} A_{kl}Y_l = \sum_{k,l=1}^{N} a_{jk}A_{kl}Y_l \\ &= \sum_{l=1}^{N} b_{jl}Y_l \qquad j = 1, \cdots, M \end{aligned} \tag{4a}$$

or, of course, in matrix notation (4b).

$$[r_{j1}]_{M,1} = [a_{jk}]_{M,N} \cdot [A_{kl}]_{N,N} \cdot [Y_{l1}]_{N,1} = [b_{jl}]_{M,N} \cdot [Y_{l1}]_{N,1} \tag{4b}$$

We can also impute the prices of N intermediate goods and commodities to the prices of the M basic resources, as follows:

$$p_k = \sum_{j=1}^{M} v_j b_{jk} \qquad k = 1, \cdots, N \tag{5a}$$

$$[p_{1k}]_{1,N} = [v_{1j}]_{1,M} \cdot [b_{jk}]_{M,N} \tag{5b}$$

To complete the system, it may be supposed that demand and supply relationships are given, a priori, by Pareto-type preference functions:

$$\text{Demand:} \quad Y_k = F_k(p_1, \cdots, p_N) \qquad k = 1, \cdots, N \tag{6}$$

$$\text{Supply:} \quad r_k = G_k(v_1, \cdots, v_M) \qquad k = 1, \cdots, M \tag{7}$$

where, of course, the p_j are functions of the v_j as in (5b).

In order to interpret the X's as physical production, it is necessary for the sake of consistency to arrange that outputs and inputs always balance, which implies that the C_{ij} must comprise *all* materials exchanges including residuals. To complete the system so that there is no net gain or loss of physical substances, it is also convenient to introduce two additional sectors, viz., an "environmental" sector whose (physical) output is X_0 and a "final consumption" sector whose output is denoted X_f. The system is then easily balanced by explicitly including flows both to and from these sectors.

To implement this further modification of the Walras-Cassel model, it is convenient to subdivide and relabel the resource category into tangible raw materials $\{r^m\}$ and services $\{r^s\}$:

$$\begin{bmatrix} r_1 \\ r_2 \\ \vdots \\ r_L \end{bmatrix} \text{ becomes } \left.\begin{bmatrix} r_1^m \\ r_2^m \\ \vdots \\ r_L^m \end{bmatrix}\right\} \text{raw materials (units)}$$

$$\begin{bmatrix} r_{L+1} \\ \vdots \\ r_M \end{bmatrix} \text{ becomes } \left.\begin{bmatrix} r_1^s \\ \vdots \\ r_p^s \end{bmatrix}\right\} \text{service (units)}$$

where, of course,

$$L + P = M \tag{8}$$

It is understood that services, while not counted in tons, can be measured in meaningful units, such as man-days, with well defined prices. Thus, we similarly relabel the price variables as follows:

$$\begin{bmatrix} V_1 \\ \vdots \\ V_L \end{bmatrix} \text{ becomes } \left.\begin{bmatrix} V_1^m \\ \vdots \\ V_L^m \end{bmatrix}\right\} \text{raw material (prices)}$$

$$\begin{bmatrix} V_{L+1} \\ \vdots \\ V_M \end{bmatrix} \text{ becomes } \left.\begin{bmatrix} V_1^s \\ \vdots \\ V_p^s \end{bmatrix}\right\} \text{labor and service (prices)}$$

The coefficients $\{a_{ij}\}$, $\{b_{ij}\}$ are similarly partitioned into two groups,

$$\text{e.g.,} \quad \begin{matrix} b_{1j} \\ \vdots \\ b_{Lj} \\ b_{L+1,j} \\ \vdots \\ b_{Mj} \end{matrix} \quad \text{becomes} \quad \begin{matrix} b_{1j}^m \\ \vdots \\ b_{Lj}^m \\ b_{1j}^s \\ \vdots \\ b_{pj}^s \end{matrix}$$

These notational changes have no effect whatever on the substance of the model, although the equations become somewhat more cumbersome. The partitioned matrix notation simplifies the restatement of the basic equations. Thus (1b) becomes (9), while (5b) becomes (10).

$$(9)\qquad M\left\{\begin{bmatrix}\vdots\\ r\\ \vdots\end{bmatrix}\right. \equiv \begin{bmatrix} r^m \\ \cdots \\ r^s\end{bmatrix}\begin{matrix}\}L\\ \\ \}P\end{matrix} = M\left\{\overbrace{\begin{bmatrix} \}L & b^m \\ & \cdots \\ \}P & b^s\end{bmatrix}}^{N}\begin{bmatrix} \\ Y \\ \\ \end{bmatrix}\right\}N$$

$$(10)\qquad \underbrace{[p_1, \cdots, p_N]}_{N} = \underbrace{\underset{L\quad P}{[v^m \vdots\, v^s]}}_{M}\underbrace{\begin{bmatrix} b^m \\ \cdots \\ b^s \\ \end{bmatrix}}_{N}\Big\}M$$

$$= [\cdots v^m \cdots]\begin{bmatrix}\vdots & b^m & \vdots \\ \vdots & \cdots & \vdots\end{bmatrix} + [\cdots v^s \cdots]\begin{bmatrix}\vdots & b^s & \vdots \\ \vdots & \cdots & \vdots\end{bmatrix}$$

The equivalent of (5a) is:

$$(11)\qquad p_k = \underbrace{\sum_{j=1}^{L} b^m_{jk} v^m_j}_{\substack{\text{prices imputed}\\ \text{to cost of raw}\\ \text{materials}}} + \underbrace{\sum_{j=1}^{P} b^s_{jk} v^s_j}_{\substack{\text{prices imputed}\\ \text{to cost of}\\ \text{services}}}$$

where $k = 1, \cdots, N$

We wish to focus attention explicitly on the flow of materials through the economy. By definition of the Leontief input coefficients (now related to materials flow), we have:

$C_{kj}X_j$ (physical) quantity transferred from k to j

$C_{jk}X_k$ quantity transferred from j to k

Hence, material flows *from* the environment to all other sectors are given by:

$$(12)\qquad \sum_{k=1}^{N} C_{0k}X_k = \sum_{j=1}^{L} r^m_j = \sum_{j=1}^{L}\sum_{k=1}^{N} a^m_{jk}X_k = \sum_{j=1}^{L}\sum_{k=1}^{N} b^m_{jk}Y_k$$

using equation (1), as modified.[18] Obviously, comparing the first and third terms,

$$(13)\qquad \underbrace{C_{0k}}_{\substack{\text{total material}\\ \text{flow (0 to } k)}} = \underbrace{\sum_{j=1}^{L} a^m_{jk}}_{\substack{\text{all raw materials}\\ (0 \text{ to } k)}}$$

Flows into and out of the environmental sector must be in balance:

$$(14)\qquad \underbrace{\sum_{k=1}^{N} C_{0k}X_k}_{\substack{\text{sum of all raw}\\ \text{material flows}}} = \underbrace{\sum_{k=1}^{N} C_{k0}X_0 + C_{f0}X_0}_{\substack{\text{sum of all return}\\ \text{(waste) flows}}}$$

Material flows to and from the final sector must also balance:

$$(15)\qquad \underbrace{\sum_{k=1}^{N} C_{kf}X_f}_{\substack{\text{sum of all}\\ \text{final goods}}} = \underbrace{\sum_{k=1}^{N} C_{fk}X_k}_{\substack{\text{sum of all}\\ \text{materials}\\ \text{recycled}}} + \underbrace{C_{f0}X_0}_{\substack{\text{waste residuals}\\ \text{(plus accumulation[19])}}}$$

[18] Ignoring, for convenience, any materials flow from the environment *directly* to the final consumption sector.

[19] For convenience, we can treat accumulation in the final sector as a return flow to the environment. In truth, structures actually *become* part of our environment, although certain disposal costs may be deferred.

Of course, by definition, X_f is the sum of the final demands:

$$(16) \qquad X_f = \sum_{j=1}^{N} Y_j$$

Substituting (16) into the left side of (15) and (2a) into the right side of (15), we obtain an expression for the waste flow in terms of final demands:

$$(17) \quad C_{f0}X_0 = \sum_{j=1}^{N}\sum_{k=1}^{N}(C_{jf} - C_{fj}A_{jk})Y_k$$

The treatment could be simplified slightly if we assumed that there is no recycling per se. Thus, in the context of the model, we could suppose that all residuals return to the environmental sector,[20] where some of them (e.g., waste paper) become "raw materials." They would then be indistinguishable from new raw materials, however, and price differentials between the two would be washed out. In principle, this is an important distinction to retain.

III. *Inclusion of Externalities*

The physical flow of materials between various intermediate (production) sectors and the final (consumption) sector tends to be accompanied by, and correlated with, a (reverse) flow of dollars.[21] However, the physical flow of materials from and back to the environment is only partly reflected by actual dollar flows, namely, land rents and payments for raw materials. There are three classes of physical exchange for which there exist no counterpart economic transactions. These are: (1) private use for production inputs of "common property" resources, notably air, streams, lakes, and the ocean; (2) private use of the assimilative capacity of the environment to "dispose of" or dilute wastes and residuals; (3) inadvertent or unwanted material inputs to productive processes—diluents and pollutants.

All these goods (or "bads") are physically transferred at zero price, not because they are not scarce relative to demand—they often are in developed economies—or because they confer no service or disservice on the user—since they demonstrably do so—but because there exist no social institutions that permit the resources in question to be "owned," and exchanged in the market.

The allocation of resources corresponding to a Pareto optimum cannot be attained without subjecting the above-mentioned nonmarket and involuntary exchanges to the moderation of a market or a surrogate thereof. In principle, the influence of a market might be simulated, to a first approximation, by introducing a set of shadow (or virtual) prices.[22] These may well be zero, where supply truly exceeds demand, or negative (i.e., costs) in some instances; they will be positive in others. The exchanges are, of course, real.

The Walras-Cassel model can be generalized to handle these effects in the following way:

1. One can introduce a set of R common-property resources or services of raw materials $\{r_1^{cp}, \cdots, r_R^{cp}\}$ as a subset of the set $\{r_j\}$; these will have corresponding virtual prices $\{v_j^{cp}\}$, which would constitute an "income" from the environment. Such resources include the atmosphere; streams, lakes, and oceans; landscape; wildlife and biological diversity; and the indispensable assimilative capacity of the environment (its ability to accept and neutralize or recycle residuals).[23]

[20] In calculating actual quantities, we would (by convention) ignore the weight of oxygen taken free from the atmosphere in combustion and return as CO^2. However, such inputs will be treated explicitly later.

[21] To be precise, the dollar flow governs and is governed by a combined flow of materials and services (value added).

[22] A similar concept exists in mechanics where the forces producing "reaction" (to balance action and reaction) are commonly described as "virtual forces."

[23] Economists have previously suggested generalization of the Walras-Cassel model to take account of public goods. One of the earliest appears to be Schles-

2. One can introduce a set of S environmental *disservices* imposed on consumers of material resources, by forcing them to accept unwanted inputs $\{r_1^u, \cdots, r_s^u\}$ (pollutants, contaminants, etc.); these disservices would have negative value, giving rise to *negative* virtual prices $\{u_j\}$.[24]

The matrix coefficients $\{a_{ij}\}$ and $\{b_{ij}\}$ can be further partitioned to take account of this additional refinement, and equations analogous to (9), (10), and (11) can be generalized in the obvious way. Equation (6) carries over unchanged, but (7) must be appropriately generalized to take account of the altered situation. Actually, (7) breaks up into several groups of equations:

$$r_k^m = G_k^m(p_1, \cdots, p_N) \qquad k = 1, \cdots, L \tag{18}$$

$$r_k^s = G_k^s(p_1, \cdots, p_N) \qquad k = 1, \cdots, P \tag{19}$$

However, as we have noted at the outset, the supplies of common-property resources and environmental services or disservices are *not* regulated directly by market prices of other goods and services. In the case of common-property resources, the supplies are simply constants fixed by nature or otherwise determined by accident or noneconomic factors.

The total value of these services performed by the environment cannot be calculated but it is suggestive to consider the situation if the natural reservoir of air, water, minerals, etc., were very much smaller, as if the earth were a submarine or "spaceship" (i.e., a vehicle with no assimilative and/or regenerative capacity). In such a case, all material resources would have to be recycled,[25] and the cost of all goods would necessarily reflect the cost of reprocessing residuals and wastes for reuse. In this case, incidentally, the ambient level of unrecovered wastes continuously circulating through the resource inventory of the system (i.e., the spaceship) would in general be nonzero because of the difficulty of 100 percent efficient waste-removal of air and water. However, although the quantitity of waste products in constant circulation may fluctuate within limits, it cannot be allowed to increase monotonically with time, which means that as much material must be recycled, on the average, as is discarded. The value of common resources plus the assimilation services performed by the environment, then, is only indirectly a function of the ambient level of untreated residuals per se, or the disutility caused thereby, which depend on the cost efficiency of the available treatment technology. Be this as it may, of course, the bill of goods produced in a spaceship economy would certainly be radically different from that we are familiar with. For this reason, no standard economic comparison between the two situations is meaningful. The measure of worth we are seeking is actually the difference between the total welfare "produced" by a spaceship economy, where 100 percent of all residuals are promptly recycled, vis-à-vis the existing welfare output on earth, where resource inventories are substantial and

inger [23]. We are indebted to Otto Eckstein for calling our attention to this key reference.

[24] The notion of introducing the possibility of negative prices in general equilibrium theory has apparently been discussed before, although we are not aware of any systematic development of the idea in the published literature. In this connection, it is worth pointing out the underlying similarity of negative prices and effluent taxes—which have been, and still are being considered as an attractive alternative to subsidies and federal standard-setting as a means of controlling air and water pollution. Such taxes would, of course, be an explicit attempt to rectify an imbalance caused by a market failure.

[25] Any consistent deviation from this 100 per cent rule implies an accumulation of waste products, on the average, which, by definition, is inconsistent with maintaining an equilibrium.

complete recycling need not be contemplated for a very long time to come.

This welfare difference might well be very large, although we possess no methodological tools for quantifying it. In any case, the resource inventory and assimilative capacity of the environment probably contribute very considerably to our standard of living.

If these environmental contributions were paid for, the overall effect on prices would presumably to be push them generally upward. However, the major *differential* effect of undervaluing the environmental contribution is that goods produced by high residual-producing processes, such as papermaking, are substantially underpriced vis-à-vis goods which involve more economical uses of basic resources. This is, however, not socially disadvantageous per se: that is, it causes no misallocation of resources unless, or until, the large resource inventory and/or the assimilative capacity of the environment are used up. When this happens, however, as it now has in most highly industrialized regions, either a market must be allowed to operate or some other form of decision rule must be introduced to permit a rational choice to be made, e.g., between curtailing or controlling the production of residuals or tolerating the effects (disservice) thereof.

It appears that the natural inventory of most common resources used as inputs (e.g., air as an input to combustion and respiratory processes) is still ample,[26] but the assimilative capacity of the environment has already been exceeded in many areas, with important external costs resulting. This suggests a compromise treatment. If an appropriate price could be charged to the producers of the residuals and used to compensate the inadvertent recipients—with the price determined by appropriate Pareto preference criteria—there would be no particular analytic purpose in keeping books on the exchange of the other environmental benefits mentioned, although they are quantitatively massive. We will, therefore, in the remainder of the discussion omit the common-property variables $\{r_j^{cp}\}$ and the corresponding virtual-price variables $\{v_j^{cp}\}$ defined previously, retaining only the terms $\{r_j^u\}$ and $\{u_{jk}\}$. The variable $\{r_j^u\}$ represents a physical quantity of the jth unwanted input. There are S such terms, by assumption, whose magnitudes are proportional to the levels of consumption of basic raw materials, subject to the existing technology. However, residuals production is not immutable: it can be increased or decreased by investment, changes in materials processing technology, raw material substitutions, and so forth.

At first glance it might seem entirely reasonable to assert that the *supplies* of unwanted residuals received will be functions of the (negative) prices (i.e., compensation) paid for them, in analogy with (7). Unfortunately, this assertion immediately introduces a theoretical difficulty, since the assumption of unique coefficients $\{a_{ij}\}$ and $\{C_{ij}\}$[27] is not consistent with the possibility of factor or process substitution or joint-production, as stated earlier. To permit such substitutions, one would have to envision a very large collection of alternative sets of coefficients: one complete set of a's and C's for each specific combination of factors and processes. Maximization of any objective function (such as GNP) would involve solving the entire system of equations as many times as there are combinations of factors and pro-

[26] Water is an exception in arid regions; in humid regions, however, water "shortages" are misnomers: they are really consequences of excessive use of watercourses as cheap means of waste disposal. But some ecologists have claimed that oxygen depletion may be a very serious long-run problem; see Cole [10].

[27] Or $\{b_{ij}\}$ and $\{A_{ij}\}$.

$$(21)\qquad [r] = \begin{bmatrix} r^m \\ r^s \\ r^u \end{bmatrix} = M\left\{\underbrace{\begin{bmatrix} a^m \\ \cdots \\ a^s \\ \cdots \\ a^u \end{bmatrix}}_{N}\begin{bmatrix} \\ \\ X \\ \\ \\ \end{bmatrix}\right\} N = \begin{bmatrix} b^m \\ \cdots \\ b^s \\ \cdots \\ b^u \end{bmatrix}\begin{bmatrix} \\ \\ Y \\ \\ \\ \end{bmatrix}$$

cesses, and picking out that set of solutions which yields the largest value. Alternatively, if the a's and C's are assumed to be continuously variable functions (of each other), the objective function could also, presumably, be parameterized. However, as long as the a's and C's are uniquely given, the supply of the kth unwanted residual is only marginally under the control of the producer, since it will be produced in strict relationship to the composition of the bill of final goods $\{Y_j\}$.

Hence, for the present model it is only correct to assume

$$(20)\qquad \overset{u}{r}_k = \overset{u}{G}_k(Y_1, \cdots, Y_N)$$

This limitation does not affect the existence of an equilibrium solution for the system of equations; it merely means that the shadow prices $\{u_{jk}\}$ which would emerge from such a solution for given coefficients $\{a_{ij}\}$, $\{b_{ij}\}$, and $\{C_{ij}\}$ might be considerably higher than the real economic optimum, since the latter could only be achieved by introducing factor and process changes.

Of course, the physical inputs are also related to the physical outputs of goods, as in (21).

Written out in full detail (21) is equivalent to:

$$(22)\quad \text{raw materials}\qquad \overset{m}{r}_k = \sum_{j=1}^{N} \overset{m}{a}_{kj} X_j = \sum_{j=1}^{N} \overset{m}{b}_{kj} Y_j \qquad k = 1, \cdots, L$$

$$(23)\quad \text{labor and technical services}\qquad \overset{s}{r}_k = \sum_{j=1}^{N} \overset{s}{a}_{kj} X_j = \sum_{j=1}^{N} \overset{s}{b}_{kj} Y_j \qquad k = 1, \cdots, P$$

$$(24)\quad \text{unwanted inputs}\qquad \overset{u}{r}_k = \sum_{j=1}^{N} \overset{u}{a}_{kj} X_j = \sum_{j=1}^{N} \overset{u}{b}_{kj} Y_j \qquad k = 1, \cdots, S$$

where, of course,

$$(25)\qquad L + P + S = M$$

The corresponding matrix equation for the prices of goods, in terms of production costs, is

$$(26)\qquad [p_1, \cdots, p_N] = [v^m \vdots\, v^s \vdots\, u]\begin{bmatrix} b^m \\ \cdots \\ b^s \\ \cdots \\ b^u \end{bmatrix}$$

Written out in the standard form, we obtain

$$(27)\qquad p_k = \underbrace{\sum_{j=1}^{L} \overset{m}{b}_{jk} \overset{m}{v}_j}_{\text{cost of raw materials}} + \underbrace{\sum_{j=1}^{P} \overset{s}{b}_{jk} \overset{s}{v}_j}_{\text{cost of labor and technical services}} + \underbrace{\sum_{j=1}^{S} \overset{u}{b}_{jk} \overset{u}{v}_j}_{\text{cost (compensation) for providing environmental disservices}} \qquad k = 1, \cdots, N$$

Evidently, the coefficients b_{jk}^{u} are empirically determined by the structure of the regional economy and its geography. It is assumed that a single overall (negative) price for each residual has meaning, even though each productive sector—and even each consumer—has his own individual utility function. Much the same assumption is conventionally made, and accepted, in the case of positive real prices.

All of the additional variables now fit into the general framework of the original Walras-Cassel analysis. Indeed, we have $2N+2M-1$ variables (r_i, Y_i, p_i, v_i) (allowing an arbitrary normalization factor for the price level) and $2N+2M-1$ independent equations.[28] If solutions exist for the Walras-Cassel system of equations, the arguments presumably continue to hold true for the generalized model. In any case, a discussion of such mathematical questions would carry us too far from our main theme.

IV. *Concluding Comments*

The limited economics literature currently available which is devoted to environmental pollution problems has generally taken a partial equilibrium view of the matter, as well as treated the pollution of particular environmental media, such as air and water, as separate problems.[29] This no doubt reflects the propensity of the theoretical literature to view externalities as exceptional and minor. Clearly, the partial equilibrium approach in particular is very convenient theoretically and empirically for it permits external damage and control cost functions to be defined for each particular case without reference to broader interrelationships and adjustments in the economy.

We have argued in this paper that the production of residuals is an inherent and general part of the production and consumption process and, moreover, that there are important trade-offs between the gaseous, liquid, or solid forms that these residuals may take. Further, we have argued that under conditions of intensive economic and population development the environmental media which can receive and assimilate residual wastes are not free goods but natural resources of great value with respect to which voluntary exchange cannot operate because of their common property characteristics. We have also noted, in passing, that the assimilative capacity of environmental media can sometimes be altered and that therefore the problem of achieving Pareto optimality reaches beyond devising appropriate shadow prices and involves the planning and execution of investments with public goods aspects.

We have exhibited a formal mathematical framework for tracing residuals flows in the economy and related it to the general equilibrium model of resources allocation, altered to accommodate recycle and containing unpriced sectors to represent the environment. This formulation, in contrast to the usual partial equilibrium treatments, implies knowledge of all preference and production functions including relations between residuals discharge and external cost and all possible factor and process substitutions. While we feel that it represents reality with greater fidelity than the usual view, it also implies a central planning problem of impossible difficulty, both from the standpoint of data collection and computation.

What, if any, help can the general interdependency approach we have outlined offer in dealing with pollution problems effectively and reasonably efficiently? A minimal contribution is its warning that partial equilibrium approaches, while more

[28] There is one redundant equation in the system, which expresses the identity between gross product and gross income for the system as a whole (sometimes called "Walras law").

[29] See, for example, the essays in Wolozin [29].

tractable, may lead to serious errors. Second, in projecting waste residuals for an economy—a regional economy would usually be the most relevant for planning and control—the inter-industry materials flow model can provide a much more conceptually satisfying and accurate tool for projecting future residuals production than the normal aggregative extrapolations.[30] The latter not only treat gaseous, liquid, and solid wastes separately, but do not take account of input-output relations and the fact that the materials account for the region must balance.

We think that in the next few years it will be possible to make improved regional projections of residuals along the lines sketched above. Undoubtedly, there will also be further progress in empirically estimating external costs associated with residuals discharge and in estimating control costs via various alternative measures. On the basis of this kind of information, a control policy can be devised. However, this approach will still be partial. Interrelations between the regional and national economy must be treated simplistically and to be manageable, the analysis must confine itself to a specific projected bill of goods.

The basic practical question which remains to be answered is whether an iterated series of partial equilibrium treatments—e.g., focusing on one industry or region at a time, *ceteris paribus*—would converge toward the general equilibrium solution, or not. We know of no theoretical test of convergence which would be applicable in this case but, in the absence of such a criterion, would be willing to admit the possible relevance of an empirical sensitivity test more or less along the following lines: take a major residuals-producing industry (such as electric power) and parametrize its cost structure in terms of emission control levels, allowing all technically feasible permutations of factor (fuel) inputs and processes. It would be a straightforward, but complicated, operations research problem to determine the minimum cost solution as a function of the assumed (negative) price of the residuals produced. If possible industry patterns—factor and process combinations—exist which would permit a high level of emission control at only a small increase in power production cost, then it might be possible to conclude that for a significant range of (negative) residuals prices the effect on power prices —and therefore on the rest of the economy —would not be great. Such a conclusion would support the convergence hypothesis. If, on the other hand, electric power prices are very sensitive to residuals prices, then one would at least have to undertake a deeper study of consumer preference functions to try to determine what residuals prices would actually be if a market mechanism existed. If people prove to have a strong antipathy to soot and sulfur dioxide, for instance, resulting in a high (negative) price for these unwanted inputs, then one would be forced to suspect that the partial equilibrium approach is probably not convergent to the general equilibrium solution and that much more elaborate forms of analysis will be required.

[30] Some efforts to implement these concepts are already underway. Walter Isard and his associates have prepared an input-output table for Philadelphia which includes coefficients representing waterborne wastes (unpublished). The recent study of waste management in the New York Metropolitan region by the Regional Plan Association took a relatively broad view of the waste residuals problem [31]. Relevant data on several industries are being gathered. Richard Frankel's not yet published study of thermal power in which the range of technical options for controlling residuals, and their costs, is being explored is notable in this regard. His and other salient studies are described in Ayres and Kneese [2].

References

1. R. U. Ayres, "Stability of Biosystems in Sea Water," Tech. Rept. No. 142, Hudson Laboratories, Columbia University, New York 1967.

2. ——— AND A. V. KNEESE, "Environmental Pollution," in U.S. Congress, Joint Economic Committee, *Federal Programs for the Development of Human Resources*, Vol. 2, Washington 1968.
3. W. J. BAUMOL, *Welfare Economics and the Theory of the State*. Cambridge 1967.
4. K. E. BOULDING, "The Economics of the Coming Spaceship Earth," in H. Jarrett, ed., *Environmental Quality in a Growing Economy*, Baltimore 1966, pp. 3–14.
5. J. H. BOYD, "Collective Facilities in Water Quality Management," Appendix to Kneese and Bower [17].
6. J. W. BUCHANAN AND WM. C. STUBBLEBINE, "Externality," *Economica*, Nov. 1962, *29*, 371–84.
7. ——— AND G. TULLOCK, "Public and Private Interaction under Reciprocal Externality," in J. Margolis, ed., *The Public Economy of Urban Communities*, Baltimore 1965, pp. 52–73.
8. G. CASSEL, *The Theory of Social Economy*. New York 1932.
9. R. H. COASE, "The Problem of Social Cost," *Jour. Law and Econ.*, Oct. 1960, *3*, 1–44.
10. L. COLE, "Can the World be Saved?" Paper presented at the 134th Meeting of the American Association for the Advancement of Science, December 27, 1967.
11. R. K. DAVIS, *The Range of Choice in Water Management*. Baltimore 1968.
12. O. A. DAVIS AND A. WHINSTON, "Externalities, Welfare, and the Theory of Games," *Jour. Pol. Econ.*, June 1962, *70*, 241–62.
13. R. DORFMAN, P. SAMUELSON AND R. M. SOLOW, *Linear Programming and Economic Analysis*. New York 1958.
14. I. FISHER, *Nature of Capital and Income*. New York 1906.
15. A. C. HARBERGER, "Monopoly and Resources Allocation," *Am. Econ. Rev.*, Proc., May 1954, *44*, 77–87.
16. K. W. KAPP, *The Social Costs of Private Enterprise*. Cambridge 1950.
17. A. V. KNEESE AND B. T. BOWER, *Managing Water Quality: Economics, Technology, Institutions*. Baltimore 1968.
18. F. H. KNIGHT, *Risk, Uncertainty, and Profit*. Boston and New York 1921.
19. H. LEIBENSTEIN, "Allocative Efficiency vs. 'X-Efficiency,'" *Am. Econ. Rev.*, June 1966, *56*, 392–415.
20. W. P. LOWRY, "The Climate of Cities," *Sci. Am.*, Aug. 1967, *217*, 15–23.
21. E. J. MISHAN, "Reflections on Recent Developments in the Concept of External Effects," *Canadian Jour. Econ. Pol. Sci.*, Feb. 1965, *31*, 1–34.
22. A. C. PIGOU, *Economics of Welfare*. London 1952.
23. K. SCHLESINGER, "Über die Produktionsgleichungen der ökonomischen Wertlehre," *Ergebnisse eines mathematischen Kolloquiums*, No. 6. Vienna, F. Denticke, 1933.
24. D. SCHWARTZMAN, "The Burden of Monopoly," *Jour. Pol. Econ.*, Dec. 1960, *68*, 627–30.
25. T. SCITOVSKY, "Two Concepts of External Economies," *Jour. Pol. Econ.*, Apr. 1954, *62*, 143–51.
26. F. SMITH, *The Economic Theory of Industrial Waste Production and Disposal*, draft of a doctoral dissertation, Northwestern Univ. 1967.
27. R. TURVEY, "On Divergencies between Social Cost and Private Cost," *Economica*, Nov. 1962, *30*, 309–13.
28. L. WALRAS, *Elements d'economie politique pure*, Jaffé translation. London 1954.
29. H. WOLOZIN, ed., *The Economics of Air Pollution*. New York 1966.
30. CONSERVATION FOUNDATION, *Implications of Rising Carbon Dioxide Content of the Atmosphere*. New York 1963.
31. REGIONAL PLAN ASSOCIATION, *Waste Management*, a Report of the Second Regional Plan. New York 1968.

Employment and Wages in Rural Egypt

By BENT HANSEN*

Egypt often appears as a classical example of overpopulation and surplus labor in development literature. Government planning has been based on the assumption that agriculture and surplus labor amounts to about 25 percent of the labor force. Yet, among those who have studied rural employment conditions in Egypt more concretely, doubts have been expressed for a long time about the realism of such assumptions, and some evidence has been presented against the hypothesis of surplus labor and zero marginal productivity of labor in agriculture. Labor requirement estimates by Mona El Tomy and Hansen [3] indicated that underemployment in agriculture is mainly a seasonal phenomenon, and production function studies by M. M. El Imam [4] and Hanaa Kheir El Dine [6] showed that rural wages were in line with marginal value productivity of labor. Both of these pieces of evidence were challenged by D. Mead [8]. Considering the shaky nature of both labor requirement and production function estimates, Hansen [1 and 2] tried to throw indirect light on the issue by showing that factor prices, wages, and rents have behaved in accordance with marginal productivity theory rather than with subsistence theory. An important contribution was finally made by A. Moheieddin [9] who, followed by R. Mabro [7], on the basis of less aggregated data, tried to reconcile the idea of surplus labor with the competitive behavior of wages.

In 1963, the International Labor Organization (*I.L.O.*), and the Institute of National Planning, Cairo, (*I.N.P.*), decided to undertake a joint, large-scale, rural employment survey. Its main purpose was to collect detailed and complete labor records from a sample of households during a full year. The survey took place from March 1964 to February 1965, with a supplementary interview survey made during a week in January of 1965. The results are now available in a series of mimeographed reports [5, IX]. These reports are not easily accessible, it would seem, and the presentation is not always comprehensible. However, the survey clears up most of the problems relating to employment in Egyptian agriculture and settles the issue of surplus labor rather definitely against the surplus labor hypothesis. By international standards it is rather unique. This article presents the main results and discusses their implications with respect to surplus labor.

The Rural Employment Survey[1]

After some pilot studies, it was decided to work with a sample of four hundred eighty households selected from forty-

* The author is professor of economics at the University of California, Berkeley. He wishes to thank Mr. Seung Chul Ahn for assistance with statistical calculations and preparation of diagrams.

[1] The survey was guided by a steering committee of experts on the various aspects of rural employment. It was administered and supervised by Dr. Mukhtar Hamza. The design of the survey was mainly the work of Dr. Salib Rofael and the *I.L.O.* expert, Dr. Ulrich Planck. They also took care of data processing and tabulation and were advised at all stages by a number of expert subcommittees. The reports were written by the various subcommittees, and a summary report was compiled by the *I.N.P.* Here we shall concentrate on Report C, *Utilization of Rural Manpower* [5, IV] and the two accompanying volumes of statistical tables [5, V], together with Report D on *Wages, Income, and Consumption* [5, VI]. The present author was a member of the steering committee and the subcommittees on Reports C and D.

eight villages in six Govornerates (*mudeyerat*). The villages were selected in pairs from four districts (*markazat*) of each Govornerate: a big village and a neighboring small village. Of the ten households selected from each village, two were non-agricultural. The eight agricultural households were to be selected so as to represent eight strata with respect to household size and size of farm. Three working members became the line of demarcation between small and big households. Agricultural households were divided into four groups: landless, i.e. with less than one-half feddan,[3] and farms with one-half to two feddan, two to five feddan, and five feddan and more. It was decided to use rotating samples so that a household was only interviewed once per four days concerning the last two days' labor; an eight-day period was chosen, therefore, as the basic unit period of tabulation. The interviewers were mainly village teachers whom the villagers might be supposed to trust and who had a good first-hand knowledge of the households and their situations. The questionnaires recorded all operations (outside household work proper) lasting more than half an hour. Household work proper was not recorded. The exclusion of small jobs implies a downward bias in the work time recorded. Another downward bias (important in particular for women and children) is related to the fact that temporary workers have not always been included.

All persons covered by the survey were classified as *men* (adult males fifteen years and above), *women* (adult females fifteen years and above), and *children* (males and females, six to fifteen). All persons six years and above were considered working members of the household where they appeared unless they were obviously incapable of doing work for reasons of health, old age, etc., or had not been recorded at all as working outside the household proper. In the sample, 72 percent of all males and 45 percent of all females were recorded as "working" [5, V, I, p. 25, Table 9D]. For males, the participation ratio of the sample is somewhat higher than that of the population census of 1960; for females it is much higher.

Although great efforts were exerted to obtain a representative sample, the results cannot be inflated to national totals. All averages presented in the following are, therefore, simple sample means, but for the six (out of a total of eighteen) Govornerates chosen, the sample is presumably representative. In retrospect, it would also seem that the choice of survey year, from March 1, 1964, to February 28, 1965, was not the best possible. Around March 1 the slack winter season ends and the very busy spring season begins. Even a slight dislocation of the seasons might thus influence the annual totals. Fortunately, the seasons are very regular in Egyptian agriculture, and for these two years there was no obvious disruption of the seasonal patterns.

The period of the study was characterized by inflationary conditions in the country. Excess demand prevailed and the official cost of living index increased by 15 percent during this period. This should presumably tend to imply good employment opportunities. Government expenditures expanded strongly and both defense and public works (such as the High Dam at Aswan) must have absorbed manpower from rural districts. It is difficult to gauge how much this has contributed to an increase in rural employment and the absorption of rural labor. The figures for working hours per person, and per year, could, perhaps, be adjusted downward by some 2–3 percent in order to allow for the effects of the inflation. In any case, this circumstance should be borne in mind.

[3] One feddan is slightly more than one acre.

TABLE 1—AVERAGE ANNUAL WORKING HOURS ACCORDING TO SEX-AGE GROUPS, TYPES OF HOUSEHOLDS, AND TYPES OF WORK

Type of household	Sex-age group	Number of hours worked annually	Percent of Annual Work Time Spent On:				
			Field work	Animal husbandry	Processing farm products	Other agricultural work	Nonagricultural work
Farmers	Men	2,280	53	21	3	13	10
	Women	869	19	63	11	3	4
	Children	1,022	49	39	3	5	4
	Total	1,642	48	30	4	10	8
Farm-laborers	Men	2,324	58	13	3	11	15
	Women	904	31	35	4	8	22
	Children	1,374	55	23	2	7	13
	Total	1,716	53	18	3	10	16
Others (nonagricultural)	Men	2,482	8	4	3	3	82
	Women	697	14	29	6	2	49
	Children	1,087	25	26	2	1	46
	Total	1,738	11	10	3	2	74

Source: [5, IV, p. 59] and [5, V, I, p. 29, Table 10A].

The Labor Records: Annual Data

A striking and unexpected feature of the labor records is the large number of hours worked per year by all categories of rural labor. Table 1 shows the annual average number of working hours recorded for men, women, and children in three broad types of households, namely, farmers (that is, cultivators, whether owners or tenants), farm laborers, and others, as well as the distribution of annual hours worked on various types of work.

For men, the annual number of hours worked is, on average, close to what corresponds to a normal eight-hour day, namely 2,280 to 2,482 hours per year. Men from other households, i.e. service workers such as barbers and shopkeepers, seem to work longer hours than both farmers and farm-laborers. It is difficult to say to what extent recorded work for these men really represents work proper; idle hours in the shop waiting for customers have been recorded as work. For this group, therefore, the figures for hours worked should be taken with a grain of salt. Women seem to work one-third of an eight-hour day, and children about half that time. Women and children belonging to farm laborer households work longer than women and children in the other categories.

In the earlier attempts to estimate labor requirements [3 and 9], the authors agreed that the requirements for plant production or field work correspond only to about 60 percent of full-time work (three hundred days). Table 1 points to almost full employment for men. The distribution in Table 1 of hours worked by type of work solves this problem. For men in farm households, field work occupied only 53 percent of the total work time. Of the time spent on agricultural work, 41 percent went to animal husbandry, processing of agricultural products, and other agricultural work. Therefore, the findings here seem to be consistent with the earlier labor

TABLE 2—AGRICULTURAL HOUSEHOLDS: HOURS WORKED ANNUALLY, TOTALLY AND OUTSIDE OWN FARM, BY SIZE OF FARM, AND FAMILY, AND AGE-SEX GROUPS

Strata		Men		Women		Children	
Size of Farm	Number of working family members	Hours worked per year	Percent of which outside own farm	Hours worked per year	Percent of which outside own farm	Hours worked per year	Percent of which outside own farm
½ to 2 feddan	3 and less	2,384	18	906	4	1,070	14
2 to 5 feddan	3 and less	2,420	4	1,112	5	1,096	9
≥5 feddan	3 and less	2,062	3	834	6	1,702	2
½ to 2 feddan	more than 3	2,190	33	1,010	9	1,122	34
2 to 5 feddan	more than 3	2,230	14	794	2	1,020	14
≥5 feddan	more than 3	2,358	5	734	1	848	11
landless[a]	3 and less	2,444	69	838	35	1,374	80
landless	more than 3	2,208	73	948	30	1,374	65

[a] With less than ½ feddan or no land at all.
Source: [5, V, I, Table 10A, p. 29, and II, Tables LRII, No. 40 to 60]. The percentages for work outside own farm are calculated on family members with at least 30 days recorded work.

requirement estimates. These earlier estimates went wrong because it was not understood what a modest part field work is of total agricultural work. For women and children, the comparison is more difficult, although the Rural Employment Survey and the earlier labor requirement estimates seem to be fully consistent.

Another remarkable feature of Table 1 is that men in farm households work about 10 percent in nonagricultural work, that is, as hired labor outside agriculture. This shows that a substantial amount of work must have been done outside the farmer's own enterprise. We shall look at this point in somewhat more detail in Table 2, because it contains crucial evidence against the notion of zero marginal productivity of labor in Egyptian agriculture.

The number of working hours for men is of the same order of magnitude in all the agricultural strata. On farms it varies between 2,062 (but see below) and 2,420 hours per year. There is no clear dependency upon farm size or family size. Nor is there any a priori reason to expect any particular relationship between hours worked per year and farm and family size. On a larger farm there may be more work to do, but, on the other hand, income is higher and the owner can afford to rely on hired labor to a larger extent. With a larger family, the need for income is higher, but so is the number of children who can work. It should be expected that women work less outside the household when the income is higher and the family is larger. There is, perhaps, such a tendency in the table, but it is by no means clear. For children, similar remarks apply, although the picture is disturbed by the very high figure for hours worked by children in big families with big farms.[3]

Concerning the time worked outside the own farm, the picture is very clear. The members of the smallest farms work a substantial part of their time outside the own farm. For men the percentages are 18 and 33, respectively, in small and big families. For children the percentages are of the same order of magnitude. For women they are much smaller, 4 and 9

[3] I have not been able to find the definite explanation of this high figure. Hours per year, per person, are calculated by dividing the tabulated total hours worked by number of individuals. However, the number of children in this group is remarkably low as compared with the number of men. I suspect, therefore, a classification mistake, where some children may have been classified as men. Indeed, the number of hours worked for men in this group is remarkably low.

TABLE 3—WORK BY NONFAMILY MEMBERS AS PERCENT OF TOTAL LABOR INPUT ON FARMS, BY FARM SIZE, FAMILY SIZE, AND SEX-AGE GROUPS

Strata		Men		Women		Children	
Size of Farm	Number of working family members	Permanent nonfamily	Permanent nonfamily plus all temporary	Permanent nonfamily	Permanent nonfamily plus all temporary	Permanent nonfamily	Permanent nonfamily plus all temporary
½ to 2 feddan	3 and less	—	7.3	—	9.4	—	26.2
2 to 5 feddan	3 and less	6.1	21.7	—	16.3	—	33.5
≥5 feddan	3 and less	19.3	52.9	—	41.1	15.4	77.8
½ to 2 feddan	more than 3	—	6.2	—	2.0	—	4.6
2 to 5 feddan	more than 3	2.1	10.9	—	13.6	—	16.2
≥5 feddan	more than 3	7.9	28.2	—	23.1	2.2	43.1

Source: [5, V, II, Tables LR II Nos. 40 to 60].

percent, respectively; this could be expected on the basis of the attitudes toward women's work. Small farmers' households thus rely heavily on income from outside. Farmers, by and large, are fully occupied, taking into account the work outside their own farms. Moreover, the work outside shows a seasonal pattern similar to that of the farms' labor input [5, IV, p. 119, Graph 10]. It seems clear, therefore, that farmers have a real choice between taking paid labor outside the farm or working on the farm. Under these circumstances, there is little reason to believe that the value of the marginal product of labor, even on the smallest farms, would be smaller than current rural wages.[4]

Work outside the own farm declines rapidly with increasing size of farm, as expected. Yet the households on smaller farms (two to five feddan), especially those farmers with large families, do substantial amounts of work outside their own farm. For the landless farm laborer households, the majority of work is, of course, done outside the own farm. However, this group may have up to one-half feddan, and we still find a considerable amount of work done on their own farm. The argument about the marginal product of labor obviously applies to this group too.

For all men in farm households, 14.2 percent of the work done was outside their own farm. Moreover Table 1 shows 10 percent of the work was nonagricultural, so that about two-thirds of the work outside the own farm is nonagricultural, including canal work for the government.

So far we have dealt exclusively with the labor performed by various types of households, that is, the output of labor by households, while nothing has been said about the input of labor on farms of various sizes. In order to arrive at the input for a certain farm, we have to deduct the household members' work outside the farm and add nonhousehold members' work (permanent or temporary) on the farm. Table 3 illuminates to what extent farms of different sizes depend on labor from nonfamily members. The distinction in this table between permanent and temporary workers is based on the number of work days which has been recorded for an individual at a given farm. If more than thirty days have been recorded, the worker is considered permanent; otherwise he is temporary. Unfortunately, no breakdown

[4] Two major forms of unpaid labor outside the own farm exist—*el mosharka*, where the peasants share work, and *el zamala*, where an exchange of labor takes place. Neither form is so important, however, that it can upset the argument of the text [5, V, I, p. 68]. For agricultural households, paid labor amounted to 96 percent of all labor outside own farms [5, V, I, p. 69].

of the temporary workers into family members and nonfamily members is available. The second column for men, women, and children includes, therefore, some work by family members. Now, it is clear that the probability that a man belonging to the family should be recorded for less than thirty days' work during a year is very small, indeed. For men, the second column must, by and large, consist of nonfamily labor. For women, the same will probably tend to be true. But for children, the chance is larger that the family's own children may have been recorded for less than thirty days work and thus appear as temporary. However, also for children, nonfamily members must dominate among the temporary workers. From the table it is seen that the percentage of child work performed by temporary workers is much larger for small families than for big families. If the family's own children dominated the temporary child work, the picture should be the opposite.

As one could expect, the use of hired labor is smaller on farms with big families than on farms with small families. Down to two feddan, farms rely heavily on hired labor and even from the very smallest farms of one-half to two feddan, there is a certain demand for hired labor. This has a bearing upon the findings of Moheieddin [9], namely, a tendency for larger farms (above ten feddan) to be concentrated in certain regions. Moheieddin [9] and Mabro [7] argue that for this reason household members from the small farms would not get in contact with labor demand where it actually exists. Labor demand seems to exist everywhere, even if there are no farms above ten feddan.

The percentage of total labor input performed by nonfamily members seems to be particularly high for children's work, and very high, indeed, for farms with small families. Women and children are hired mainly for cotton picking, weeding, and similar work, and to judge from the seasonal data, a typical pattern of small-farm behavior is to hire women and children for this kind of work at the same time as the men of the farm take hired labor outside the farm. The reason for this is obvious: a man can usually earn more outside the farm than he has to pay women and children for work which they can do as well or better than he can.

TABLE 4—AVERAGE ANNUAL WORK TIME PE WORKING PERSON BY REGIONS AND SEX-AGE GROUPS (hours per year)

Govornerate	Men	Women	Children
Delta			
Beheira	2,579	1,134	1,616
Gharbiya	2,042	884	889
Menufiya	2,531	1,391	1,462
Upper Egypt			
Assiout	2,141	562	698
Quenna	2,244	755	1,180
Fayoum	2,330	459	1,064

Source: [5, IV, p. 90].

Finally, a distribution by regions is of interest because Upper Egypt should, a priori, be expected to offer fewer employment opportunities than the Delta. Substantial areas in Upper Egypt were still basin irrigated in 1964–65, with only one crop per year. Table 4 shows the average annual hours of work by Govornerates and sex-age groups.

The table confirms Mead's result [8] that employment is at a lower level in Upper Egypt than in the Delta. He argued that at least in Upper Egypt there was heavy underemployment. But the lower employment in Upper Egypt is concentrated mainly among women and children. Taking simple averages of the Governerate averages for the two areas, the annual number of working hours for men in the Delta is slightly below 2,400 hours, while for Upper Egypt the corresponding figure is slightly below 2,250 hours. The difference is only about 6 percent. Women,

however, work much less in Upper Egypt than in the Delta, slightly below 500 hours as against almost 1,150 hours. For children the difference is also substantial, slightly below 1,000 hours as against almost 1,325 hours. Thus, it would seem that woman and child labor serves as a buffer, with the men working about full time and the women and children doing what has to be done beyond that. The low figures for women in Upper Egypt may also be related to the much more conservative attitudes and traditions in that area with respect to women's activities,[5] although it is difficult to say here what is cause and effect. Attitudes and traditions have a wonderful ability to adjust themselves to the necessities of life. Similar attitude differences do not exist for child labor. We should also point to the much hotter climate of Upper Egypt which may necessitate shorter working hours than in the Delta.

The Labor Records: Seasonal Fluctuations

The seasonal employment pattern revealed by the labor records confirms, by and large, the earlier labor requirement estimates insofar as there seem to be two clearly distinguishable seasonal peaks for agricultural households; one in May–June and one in September. The first peak is related to the wheat and maize harvest; the second to the cotton harvest. The slack season occurs from the middle of October to the middle of February. The difference between seasonal peaks and troughs is much smaller, however, than that found in the labor requirement studies. The reason is that types of work not included in the labor requirement studies either have little seasonality (animal husbandry) or may be counterseasonal (canal work). For adult male farmers and family farmhands the peak is around June 1, with slightly above fifty-seven hours work per eight-day period; while during the slack period, average work time for this category remains above forty hours (disregarding January which that year coincided with the fasting month, Ramadan, and where work time drops further to 36 hours). For farm laborers, the seasonal spread is almost the same, from sixty-one hours per eight-day period in mid-September to around forty-four hours (thirty-seven during the Ramadan). The seasonality of craftsmen labor is quite similar to that of farmers and farmhands, while salespeople have quite a different seasonal pattern with work time increasing from May to September-October and thereafter falling sharply. This may be related to the pattern of cash earnings in agriculture, but may also simply reflect the fact that during the hot seasons, sales take place, to a larger extent, during the cooler nights. For men, women, and children, the seasonal employment pattern appears in Figures 1, 2, and 3. Hours worked by women and children have somewhat stronger seasonal fluctuations than men. Finally, there does not seem to be any systematic difference between the seasonal fluctuations in the Delta and Upper Egypt. Everywhere men have long working hours during the busy seasons and are, at least during these seasons, fully employed.

Unemployment

In the theoretical sense of people willing to work at going market wages, supply of labor could not be measured directly. Hypothetical questions of this type do not make sense to illiterate villagers. The true supply of labor is not measured by the labor record of working persons nor by actual employment. An attempt to measure involuntary unemployment directly was carried out in the larger interview sur-

[5] These attitudes should probably show themselves also in lower participation ratios for women in Upper Egypt. It is not possible from the published data to see whether this really was the case.

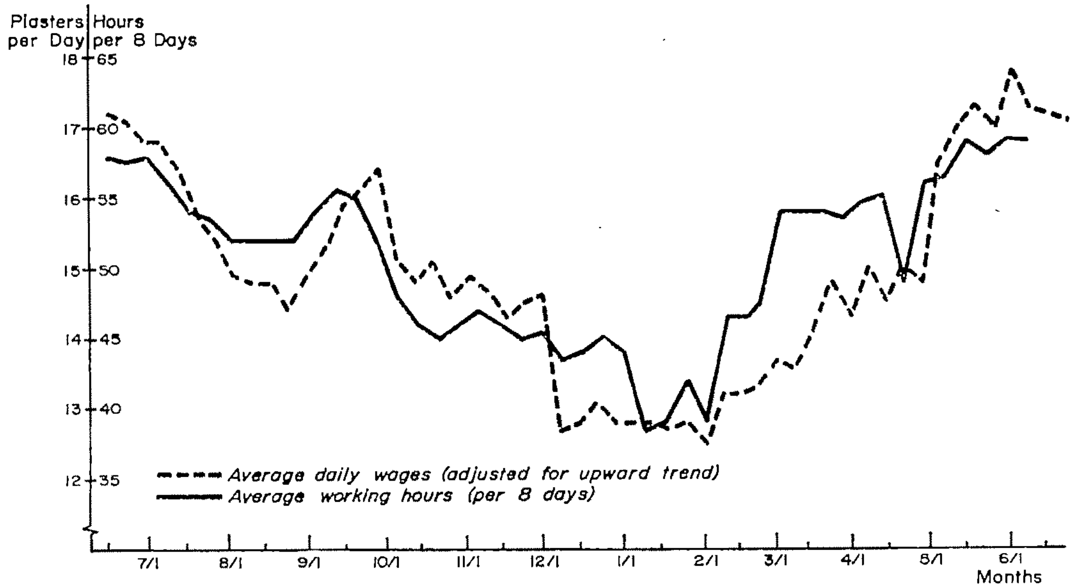


FIGURE 1. EMPLOYMENT BY EIGHT-DAY PERIODS AND DAILY WAGES—MEN.

vey. This survey was undertaken in January and thus should tend to show the maximum seasonal unemployment. To judge from this survey [5, IV, pp. 157–61], unemployment proper seems to have been about 8–9 percent of the "labor force" at the seasonal trough.[6] Since this particular

[6] This result is a little higher than the unemployment percentages found in rural districts by the labor force surveys of the Ministry of Labor.

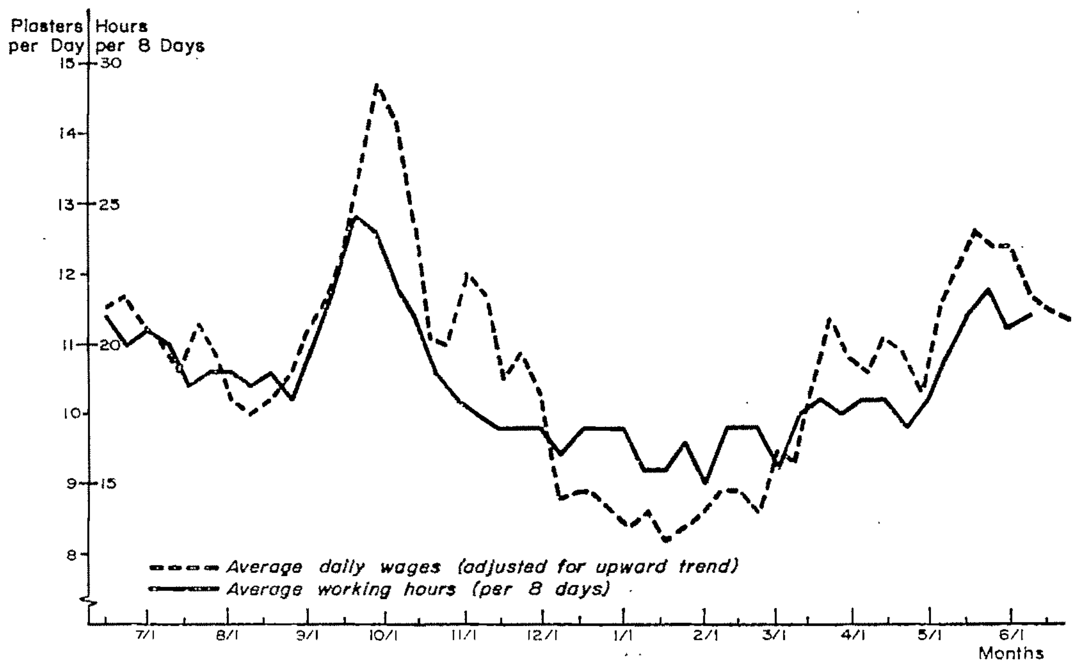


FIGURE 2. EMPLOYMENT BY EIGHT-DAY PERIODS AND DAILY WAGES—WOMEN.

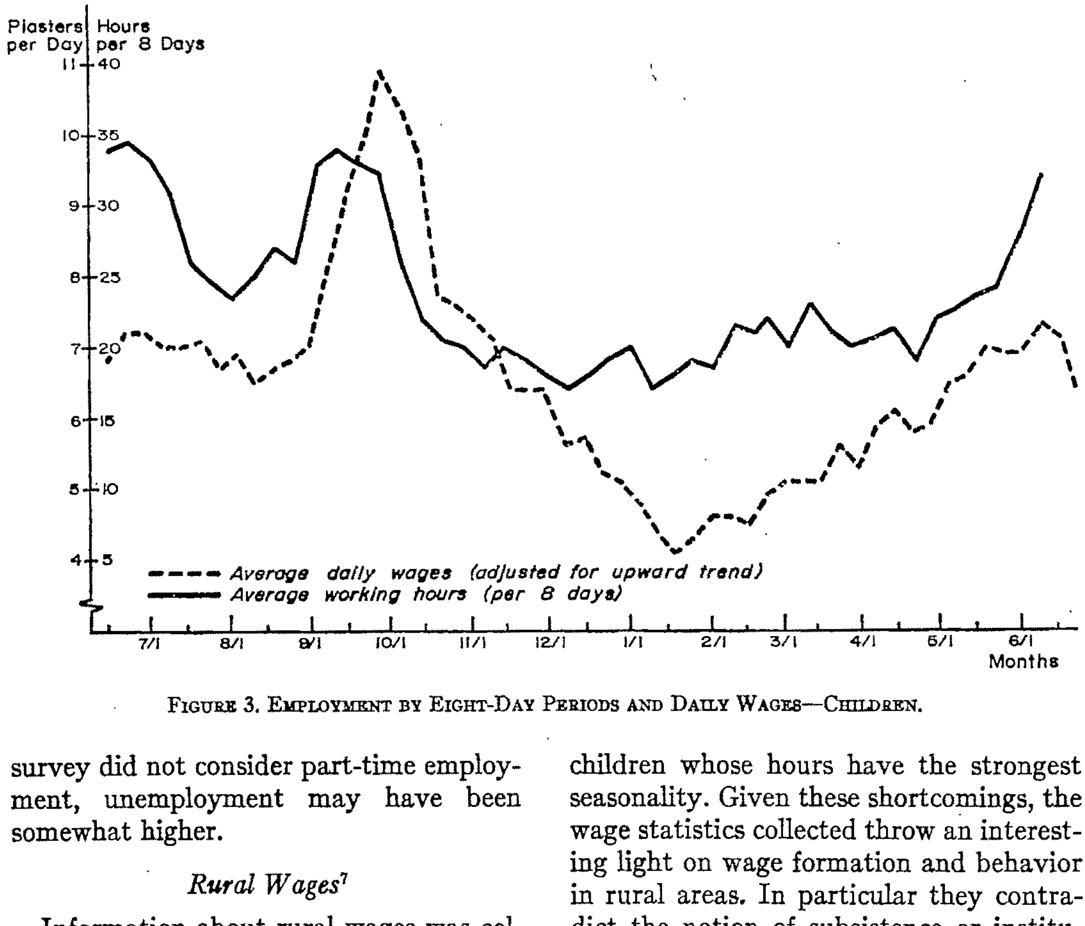


FIGURE 3. EMPLOYMENT BY EIGHT-DAY PERIODS AND DAILY WAGES—CHILDREN.

survey did not consider part-time employment, unemployment may have been somewhat higher.

Rural Wages[7]

Information about rural wages was collected together with the labor records. The laborers and the employers were not directly asked about wages received or paid. Rightly or wrongly, this was thought to be a delicate question which might call forth negative reactions. The village *sheikh* or *oumda* was asked weekly how much was paid per day to men, women, and children for various kinds of work. The wage records were not complete. From some villages, information is missing entirely. For other villages information is missing for the slack season where, during certain weeks, there may have been no wages to record. In such cases it is somewhat difficult to talk about an annual average. This problem appears mainly for women and children whose hours have the strongest seasonality. Given these shortcomings, the wage statistics collected throw an interesting light on wage formation and behavior in rural areas. In particular they contradict the notion of subsistence or institutional wages.

If rural wages were strongly influenced by a subsistence level which acted as a floor for actual market wages, we should expect the distribution of villages with respect to average wages to be truncated and squeezed up against the subsistence minimum, as illustrated in Figure 4.

The distribution actually found looks fairly normal for men, women and children as shown in Table 5. The figures for all villages (with records) may be biased for women and children due to incomplete records (see below). The right-hand part of the table therefore shows the distribution for the smaller number of villages with satisfactory records (fifty weeks and more); these figures, however, may be biased in another way since they contain

[7] The material in this section is based on [5, VI], which the present author prepared.

no villages from Upper Egypt, apart from Fayoum. The only sign of concentration at the lower end of the distributions are for men, where we find six villages at ten piasters per day.[8] These six villages are all situated in Fayoum, a kind of oasis separated from the Nile Valley by some ten miles of desert. Should this represent a subsistence level, the implication would be that the whole Nile Valley proper is substantially above the subsistence level, and that for the valley itself we have to look for other explanations of rural wages. It may be pointed out also that during the

[8] According to the official exchange rate, one piaster is equal to 2.25 U.S. cents.

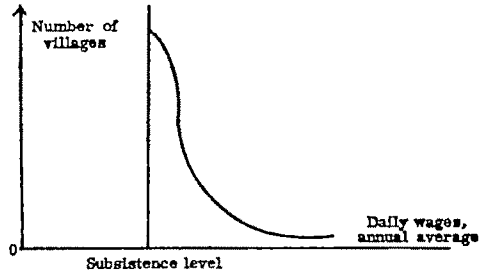


FIGURE 4

slack season daily wages in Fayoum are considerably below the average of 10 piasters.

Table 5, in addition, gives information about the average daily wages by Govornerate and for all villages included in the

TABLE 5—DISTRIBUTION OF VILLAGES BY SEX-AGE GROUPS AND SIZE OF DAILY WAGES
(annual averages)

Piasters Per Day	Number of Villages with Annual Average of Daily Wages					
	All Villages with Records			Villages with Records for 50 Weeks and Over		
	Men	Women	Children	Men	Women	Children
4–5.9		3	8			4
6–7.9		4	4			2
8–9.9		4	16		4	6
10–11.9	6	8	13	6	6	7
12–13.9	1	6	5	1	3	
14–15.9	4	10		2	2	
16–17.9	12	3		7	1	
18–19.9	11			7		
20–21.9	8			5		
22–23.9	3			2		
24–25.9	2			1		
26–27.9	1					
Total Number of Villages	48	38	46	31	16	19
Govornerate	Piasters Per Day					
Beheira	18.5	12.5	10.0	18.3	11.8	9.2
Gharbia	18.7	11.6	9.6	18.8	11.5	9.6
Menoufia	21.4	13.0	10.0	21.0	12.8	10.0
Assiout	17.6	11.7	9.8	17.6	10.6	n.a.
Kenna	20.4	11.9	9.0	19.8	n.a.	n.a.
Fayoum	11.7	6.1	5.1	11.7	n.a.	4.9
Total Village Average	18.0	11.7	8.7	17.3	11.9	8.6

Source: [5, VI, p. 7].

TABLE 6—DISTRIBUTION OF VILLAGES ACCORDING TO WAGE DIFFERENTIALS FOR WOMEN AND CHILDREN

Percent of Men's Wages	Number of Villages			
	All Villages with Records		Villages with Records for 50 Weeks and Over	
	Women	Children	Women	Children
35–40		1		
40–45		8		3
45–50	1	15		8
50–55	2	7		2
55–60	11	6	3	3
60–65	6	6	3	1
65–70	7	2	4	
70–75	3			
75–80	5		2	
80–85	2	1	1	
85–90	1			
Total Number of Villages	38	46	13	17

Source: [5, VI, p. 13].

sample. There are substantial differences between Govornerates, with Menoufia (near Cairo) at the top, and Fayoum at the bottom. The average for all villages is 18.0 piasters for men, and an institutionalist might point out that this is exactly the statutory minimum wage for male rural workers. However, the distribution in the upper and lower parts of the table shows that this is a lucky statistical coincidence. Historically, the development has been that the statutory minimum of 18 piasters for men was introduced in 1952 as part of the Land Reform of that year, followed by a government decision to make no attempt to enforce the minimum wage in rural areas.[9] Average daily wages stayed at about 12 piasters until the beginning of the sixties. Thanks to increased productivity in agriculture and inflation, rural wages have moved rapidly upwards since then. In 1964–65 their average passed the level of the statutory minimum, and today the average is probably somewhat higher.

The average wage for women is about two-thirds, and for children half of that paid men. These wage differentials were already observed in the thirties (and even before World War I) and seem to have been fairly constant over a long period. They have sometimes been interpreted as an institutional law for wage differentials. Once more we are confronted with deceptive statistical averages, however. Table 6 shows the distribution of villages according to differentials between annual averages of daily wages. The wage differentials vary widely between villages. For women the wages vary between 45–50 and 85–90 percent of men's wages, and for children the variation is greater. Even if we move to the figures of the right-hand half of the table, the variation remains large. It should be added that the variations of the differentials are also sizable within Govornerates. The proportions of two-thirds and one-half for women's and children's wages thus seem to be statistical averages without any deeper meaning.

Table 7 shows that stable, institutional, wage differentials do not exist even within the individual villages. It shows the distribution of all weeks with wage records, according to wage differentials, for women and children, for three selected typical villages with large, medium, and small variations in the differentials within the year. In Kilishan, women's and children's wages are equal to men's at the peak demand for woman and child labor (the cotton harvest), but drop to 30–39 and 20–29 percent of men's wages in the slack season. Even the village with the most stable differentials, El Nigila, has appreciable variation during the year. There is no doubt that the variation of the differentials is related to the seasonal fluctua-

[9] The minimum wages have, in principle, been applied by the government itself in rural areas. But it is difficult to say how this has worked in practice when contractors have been in charge of the works.

TABLE 7—DISTRIBUTION OF WEEKLY RECORDS ACCORDING TO WAGE DIFFERENTIALS FOR WOMEN AND CHILDREN IN THREE SELECTED VILLAGES

Percent of Men's Wages	Number of Weekly Records					
	Kilishan		El Haddad		El Nigila	
	Women	Children	Women	Children	Women	Children
20– 29.9		2				
30– 39.9	4	9				
40– 49.9	11	9	4	4		29
50– 59.9	15	20	9	17		13
60– 69.9	6	6	25	17	38	10
70– 79.9	7		5	5	12	
80– 89.9	3	2				
90– 99.9					2	
100–109.9	5	5	6	6		
Total Number of Weekly Records	51	53	49	49	52	52
Average Percent	58	49	63	61	67	48

Source: [5, VI, p. 14].

tions of the demand for various types of labor.

In Figures 1, 2, and 3 we bring together wage and employment series for men, women, and children. This comparison reveals a close association between seasonal employment and wage fluctuations. Some problems arose in constructing the series. They are explained in detail in [5, VI, pp. 21 ff.]. The employment series are derived from the total sample, while the wages are averages for villages with a sufficient number of weekly records. This comparison creates a problem because wage records for women and children are mainly from the Delta and Fayoum. The employment series are on an eight-day period basis; the wage series are on a weekly basis.[10] Moreover, the wage series had to be adjusted for a strong upward trend due to the inflation. The original employment series covered the period from March 1, 1964 to February 28, 1965, whereas the wages series covered the period from June 15, 1964 to June 28, 1965 (for men). Employment series for the period from June 22, 1964 to June 21, 1965 were constructed, therefore, by assuming that the figures for the period, March 1 to June 22, 1964 could be applied to the same period of the following year. This presumes a high stability of the seasons and of employment conditions, and that there had been no change in the quality of the labor records during the year. We think that these conditions were roughly fulfilled, but the assumptions are difficult to test, partly because the employment series cover exactly one year with no overlap.

The level at the end of the year for women and children was essentially the same as at the beginning of the year. There was a substantial difference for men, the work time being about six hours less for the last recorded eight-day period

[10] Laborers are usually hired on a daily basis and there does not appear to be any seasonal fluctuation in the length of a laborer's daily working hours. Hourly wages, therefore, show the same seasonal pattern as daily wages. See [5, VI, p. 29 and Graph 5]. This point is of course crucial for the interpretation of the relations between daily wages and employment.

than for the first. This implies that the average work time should have increased by six hours from an eight-day period including the last eight days of February to an eight-day period including the first eight days of March. A priori, this cannot be ruled out as entirely unlikely because at this time of the year there is a very strong increase in agricultural activities related to cotton sowing, and this is mainly men's work. It is impossible to exclude the possibility of bias although it is difficult to see why there should be such a general bias for men, and not for women and children who are actually recorded for a slightly higher work time in the last than in the first period.[11]

With these reservations, Figures 1, 2, and 3 point to a close association between the employment and the wage series for men, women, and children. If one identifies employment with demand for labor, and assumes that, from a seasonal point of view, wage and employment fluctuations are the result of a shifting demand curve and a fixed supply curve, the graphs then clearly support the hypothesis of demand and supply determined wages. The seasonal fluctuations of the daily wages are very strong. The highest daily wage is about 50 percent above the lowest for men; the corresponding figures are about 100 and 175 percent, for women and children, respectively.

The adjustments of the wage series of Figures 1, 2, and 3 for inflation were made by comparing wage averages for two corresponding two-week periods in mid-June in 1964 and 1965, and a linear inflationary increase during the year was then assumed. Another method of adjusting for the inflationary increase is to regress wages on employment and time. Here we meet the obstacle that the wage series contained fifty-two observations of weekly averages, whereas the employment series contained forty-six observations of average hours worked per eight-day period. We preferred to interpolate on the employment series (since this was the smoothest one) to obtain fifty-two observations for seven-day periods. The results were:

$$\textit{Men} \quad W_t = \underset{}{5.630} + \underset{(0.023)}{0.219}\, H_t + \underset{(0.008)}{0.073}\, T, \quad R = 0.88$$

$$\textit{Women} \quad W_t = -1.269 + \underset{(0.058)}{0.721}\, H_t + \underset{(0.007)}{0.052}\, T, \quad R = 0.88$$

$$\textit{Children} \quad W_t = 3.431 + \underset{(0.040)}{0.173}\, H_t + \underset{(0.013)}{0.060}\, T, \quad R = 0.61$$

where W denotes average daily wages, H average number of hours worked, and T (running from 0 to 52) is time. All coefficients are highly significant, and the trend corresponds to an increase of money wages by about 30 percent during the year, which is about the same found by the simpler method underlying Figures 1, 2, and 3. The correlation coefficients for children are somewhat lower than for men and women. Examination of Figure 3 shows a lag between employment and wages. This lag is most certainly related to difference of coverage of the wage and employment series mentioned above, the wage series for children covering the Delta and Fayoum only, whereas the employment series also covers the rest of Upper Egypt. Children's employment is closely related to the cotton crop, and all major operations related to cotton are done some weeks earlier in Upper Egypt than in the Delta. We should, for that reason, expect a lag of some few weeks between our wage and employment series for children. We found the highest correlation coefficient in using a two-week lag:

[11] A break-down on various crop rotation systems indicates that for sugar cane farms there is certainly a downward bias during the year. The year ends with only about half the average number of hours worked per feddan on these farms than at the beginning of the year. This cannot be accounted for by a seasonal increase around March. This strata is relatively unimportant, however, and mainly located at Quenna.

$$\textit{Children}\ W_t = \underset{}{0.781} + \underset{(0.040)}{0.281}\, H_{t-2} + \underset{(0.013)}{0.085}\, T, \quad R = 0.74$$

Presumably, the correlations would have been stronger if it had been possible to use wage and employment series relating to identical villages. The published data did not permit that. In any case, both figures and regressions seem to leave little doubt that the very substantial seasonal fluctuations in wages are governed by fluctuations in labor demand.[12]

Conclusions

In order to discuss rural employment in Egypt, and presumably in many other underdeveloped countries, a clear distinction has to be made between adult males, adult females, and children. The level of employment for these three categories and the functions they perform differ widely. Moreover, the social and developmental implications of men's, women's, and children's work are radically different.

The Rural Employment Survey of the *I.L.O.* and *I.N.P.*, Cairo, shows, first that the level of employment in rural areas is *much* higher than had earlier been thought. By ignoring nonfield work and nonagricultural work, earlier labor requirement calculations systematically and grossly underestimated actual employment.

Male adults seem, by and large, to be fully employed with long working hours during spring and summer, and with some underemployment from October to February. The seasonal variations appear to be much smaller than those shown by earlier labor requirement calculations. There does not seem to be any great difference in the level of employment for male adults in the various sample strata and regions. The differences in the employment possibilities have consequences mainly for the employment of women and children. Small farmers are brought to a high level of employment by the substantial opportunities for obtaining employment outside their own farm, on other farms, and outside agriculture.

Female adults seem, on the average, to work one-third of the time outside the household farm. Taking into account household work in a rather mechanical and arbitrary way, it appears that female adults in rural areas are slightly underemployed. The differences between sample strata and regions are much larger than for male adults. Female work outside the household is particularly low in Upper Egypt (South of Cairo). This may be due to lack of employment but may well reflect more conservative traditions. Female adults on small farms and in landless laborer families in the Delta more often than not are heavily overemployed (including household work), whereas on bigger farms and in Upper Egypt they seem to be able to live a more leisurely life.

Children of both sexes between six and fifteen work, on the average, slightly more than women outside the households. The variations between strata and regions are somewhat smaller for children than women; but Upper Egypt seems also to offer (or require) less employment for children than the Delta.

Wages generally appear to be highly flexible and react strongly and rapidly to changes in demand. There are large differences between villages and regions as well as between seasons, and for various types of work. Wage differentials reflect differences between demand and supply, and there is nothing to indicate that the wage level should be governed by a subsistence minimum, or by institutional factors.

The Rural Employment Survey pictures an active labor market in which even very

[12] The reader will notice that the estimated relations are *not* the Phillips curve relationship. There is nothing to indicate that the rate of change of money wages should be determined by the level of employment. The simple correlation between rate of wage change and employment level is low and insignificant.

small farms participate, both on the demand and the supply side. The farm household members take paid work outside their own farm (on other farms or outside agriculture) and hire laborers for the farm, sometimes simultaneously. The two phenomena can be correlated to the size of the farm, the former negatively and the latter positively. This active, pervasive labor market, which seems to exist everywhere, is difficult to reconcile with the idea of surplus labor as a general phenomenon with zero productivity of labor.

No information is available on the supply of labor (in the sense of labor offered at the current wages), but with the employment actually enjoyed by men it is inconceivable that men should want to work substantially more than they do at current wages during the busy half of the year. Apart from seasonalities, the only feature of rural employment which points in the direction of some elasticity of supply of labor is the employment of women and children which apparently adapts itself to employment opportunities. The fact that both women and children, particularly the former, do more outside work in the Delta than in Upper Egypt might be considered a proof of underemployment in Upper Egypt. There is a discernible tendency for woman and child wages to be relatively low in Upper Egypt, and this might suffice to establish an equilibrium at a lower level of employment. More likely, however, it is the conservative traditions of the South with respect to woman labor which is the decisive factor: it keeps the supply of woman labor down and thus prevents woman wages from being particularly low in spite of a relatively weak demand for such labor.[13] We are involved here in the tricky problem of woman and child supply of labor from family units where the men are already employed. Theory has little to tell us about this, and there have been few empirical studies of labor supply at very low income levels. But it is tempting to see here a (long term) mechanism similar to the well known (short term) adaptation of labor supply to labor demand in the United States and other developed countries.

We meet here an elasticity of labor supply very different from the unlimited supplies of labor which Lewis and others had in mind. The policy implications are radically different. Whatever the relationship between supply proper and actual employment of women and children, there can be no doubt that when we turn to social norms and development possibilities, the women and children do not represent a reserve of labor available for industrialization. Children between six and fifteen who are employed between one-third and one-half their time are certainly *overemployed* by any reasonable, modern, social standard. Thinking in terms of development programs, child labor is simply an obstacle which has to be abolished. A sensible development program must presumably include efforts to have children attend school regularly, and this must imply a reduction of child labor in agriculture and elsewhere. In the sample of the Rural Employment Survey, children worked about 15 percent of the total number of hours worked; women account for another 15 percent. If the child labor in Egyptian agriculture had to be done by the women, they would in all probabilities become heavily employed; whether this would also imply "overemployment" in relation to "supply" or to social standards

[13] All this may be history now. The increase in perennial irrigation in Upper Egypt which will become one of the major benefits from the High Dam at Aswan will also increase the demand for labor. Conditions of agricultural production are changing and should become, before the end of this decade, quite similar to those in the Delta. It will be interesting to see if traditions will give way with respect to woman labor as they did in the Delta, when employment and income possibilities there increased many decades ago.

is more difficult to say. Some child labor is naturally compatible with good educational programs (e.g., holidays), but satisfactory education for the village children would presumably, on balance and *ceteris paribus*, imply an increased demand for adult hired labor in agriculture.

REFERENCES

[1] B. HANSEN, "Marginal Productivity Wage Theory and Subsistence Wage Theory in Egyptian Agriculture," *Jour. Develop. Stud.*, July 1966, *2*.

[2] ———. "The Distributive Shares in Egyptian Agriculture, 1897–1961," *Internat. Econ. Rev.*, June 1968, *9*, 175–94.

[3] ———, AND MONA EL TOMY. "The Seasonal Employment Profile in Egyptian Agriculture," *Jour. of Devel. Stud.*, July 1965, *1*, 399–409.

[4] M. M. EL IMAM, "Some Remarks on the Labour-Force in the Egyptian Region," Memo No. 68. Cairo: *I.N.P.* (Oct. 1961; 2 ed, Apr. 1966).

[5] INSTITUTE OF NATIONAL PLANNING, *Employment Problems in Rural Areas*, U.A.R.

(I) *Report on Methodology, Field Work, and Documentation*, prepared by Salib-Rofael and Ibrahim Metaweh, Cairo, Dec. 1965.

(II) *Report A on Impact of Population Trends in UAR on Rural Employment*, prepared by Salib Rofael, Abd El-Khalik Zikry, and Kamal Mazloum, Cairo, Dec. 1965.

(III) *Report B on Migration in the UAR*, prepared by Ibrahim E. Metaweh and Mukhtar Hamza Cairo, Dec. 1965.

(IV) *Report C on Utilization of Rural Manpower and Measurement of its Underemployment*, prepared by M. Mohiey El-Din Nazrat, Zaki Shabana, Salib Rofael, U. Planck, and Q. Qayoum, Cairo, Aug. 1966.

(V) *Statistical Tables, The Labour Record Sample Survey*, Part I and II, Cairo, Dec. 1965.

(VI) *Report D on Wages, Income, and Consumption in Rural Areas*, prepared by Bent Hansen, Atef Sedki, and Yousry Moustafa, Cairo, Dec. 1965.

(VII) *Report E on Impact of National Development Projects*, prepared by Abdelmonem Shawky *et al.*, Cairo, Dec. 1965.

(VIII) *Report F on Impact of Social and Educational Policy*, prepared by Mukhtar Hamza, Abdelmonem Shawky, Badr El Din Aly, Mohammed Kazem, and Seif El Din Fahmy, Cairo, Dec. 1965.

(IX) *Tabulation Concerning the Questionnaire of the Interviews Survey*, Part I and II, prepared by Atef Khalifa, Nabil El Khorazaty, and Khader Abu Kora, Cairo, Dec. 1965.

(X) *Final Report on Employment Problems in Rural Areas, UAR* Cairo, n.a.

[6] H. KHEIR EL DINE, "The Cotton Production Function in the U.A.R. and Its Relation to Technical Progress and to Disguised Unemployment," I.N.P. Memo No. 370, Cairo 1963.

[7] R. MABRO, "Industrial Growth, Agricultural Underemployment and the Lewis Model. The Egyptian Case. 1937–1965." *Jour. Devel. Stud.*, July 1967, *3*, 322–51.

[8] D. MEAD, *Growth and Structural Change in the Egyptian Economy*. Homewood, Ill.: Irwin, 1967.

[9] AMR MOHEIEDDIN, *Agricultural Investment and Employment in Egypt Since 1935*, unpublished Ph.D. thesis, London University 1966.

The Interlocking Directorate

By PETER C. DOOLEY*

Early in this century interlocking directorates were publically attacked from many quarters. Louis Brandeis, one of the most outspoken critics and one of President Wilson's chief advisers on the trust problem, described interlocking directorates with the following words:

> The practice of interlocking directorates is the root of many evils. It offends laws human and divine. Applied to rival corporations, it tends to the suppression of competition and to violation of the Sherman law. Applied to corporations which deal with each other, it tends to disloyalty and to violation of the fundamental law that no man can serve two masters. In either event it leads to inefficiency; for it removes incentive and destroys soundness of judgement. It is undemocratic, for it rejects the platform: "A fair field and no favors,"—substituting the pull of privilege for the push of manhood [4, p. 51].

The Clayton Act of 1914 prohibited interlocking directorates among competing corporations, but it did not condemn the practice in general.

In the 1930's the National Resources Committee found that 225 of the 250 largest U. S. corporations had at least one director who sat on the board of at least one other of the largest corporations. It further discovered that 106 of these corporations belonged to " . . . eight more or less clearly defined interest groups [24, p. 161]." These findings have provoked repeated studies and comment by the government [21] [22] [23], by economists [7] [11], by sociologists [9] [13] and others [5] [12]. Paul Sweezy, who helped prepare the National Resources Committee study, has recently stated that the network of interlocking directorates has changed since the 1930's and that the concept of the interest group is now obsolete [1, pp. 17–20].

This paper investigates the nature of interlocking directorates and interest groups for 1965, compares the 1935 findings of the National Resources Committee with the current situation, and examines several reasons for the existence of interlocking directorates.

I. *Interlocks:* 1935 *and* 1965

The National Resources Committee studied the 200 largest nonfinancial corporations and the 50 largest financial corporations ranked by assets.[1] This paper uses a similar group of corporations thirty years later. The list of corporations was taken from the Fortune Directory [15]. Of course, membership in the top 250 corporations changed substantially over this period. Only 140 of the largest corporations in 1965 can readily be identified on the 1935 list, though the actual overlapping is greater than this due to mergers and reorganizations. The 200 largest nonfinancial corporations are here further subdivided into 115 industrial, 10 merchandising, 25 transportation, and 50 public utility corporations. The financial group includes 32 banks and 18 life insurance companies. The list of directors for these companies was obtained from Standard and Poor's *Register* for 1965 [20].[2]

* The author is an assistant professor at the University of Saskatchewan. He wishes to thank Anne Moore Dooley and Irene McClean who performed much of the clerical work in the study and an anonymous referee for his helpful suggestions.

[1] Most studies have measured size by assets [2] [7] [8].

[2] Moody's *Manuals* [16] [17] [18] [19] and the House Antitrust Subcommittee's *Interlocks in Corporate Management* [22] were used as cross checks when additional information was needed to identify individual directors.

TABLE 1—DISTRIBUTION OF DIRECTORSHIPS

Number of directorships held by one man	1935[a]		1965	
	Number of men	Number of directorships	Number of men	Number of directorships
1	2234	2234	2603	2603
2	303	606	372	744
3	102	306	123	369
4	48	192	49	196
5	19	95	13	65
6	6	36	5	30
7	6	42		
8	3	24		
9	1	9		
Total	2722	3544	3165	4007

[a] U.S. National Resources Committee, *The Structure of the American Economy.* Washington 1939. P. 158.

TABLE 2—INTERLOCKING DIRECTORATES AMONG THE 250 LARGEST U.S. CORPORATIONS IN 1935 AND 1965

	1935[a]		1965	
	Number of companies	Number of companies interlocked	Number of companies	Number of companies interlocked
Industrial and merchandising	107	91	125	113
Transportation	39	38	25	25
Utilities	54	46	50	46
All nonfinancial	200	175	200	184
All financial	50	50	50	49
Total	250	225	250	233

[a] U.S. National Resources Committee, *The Structure of the American Economy.* Washington 1939. P. 159.

The frequency of interlocks in 1935 and 1965 is remarkably similar. In 1965 a total of 4007 directorships were held by 3165 men. While most of these directors sat on a single board, 562 sat on two or more boards. Five men held six directorships each. In all, 1404 directorships were held by multiple directors. In 1935 there were somewhat fewer directors and directorships, though the distribution of multiple directorships was slightly more concentrated than in 1965 (Table 1)[3].

More of the top 250 corporations were interlocked in 1965 than in 1935. In the earlier period 25 corporations did not interlock at all, while only 17 were not interlocked in the later period (Table 2). In 1965 these noninterlocked companies consisted of eleven industrials, four utilities, one merchandiser, and one life insurance company.

The number of interlocks per corporation was unevenly distributed. Three companies in 1965 had directors who held 40 or more outside directorships, while 19 additional companies interlocked 20 or more times with other corporations among the top 250. Financial companies interlocked more frequently than did nonfinancial companies. In 1965 banks interlocked an average of 16.1 times compared to 9.9 times for all 250 corporations (Table 3).

[3] The Gini index of concentration is .20 for 1935 and .18 for 1965, where .00 indicates perfect equality.

II. *Why Do Interlocks Occur?*

The institution of the interlocking directorate has continued to exist since the early days of corporate capitalism. This is of some interest in itself, because it is doubtful that it would have survived without serving some material purpose. The critical question is what purpose (or purposes) does it serve.

Like many social phenomena, the interlocking directorate is shaped by a multitude of tangible and intangible forces. Yet interlocks occur with sufficient order to permit an empirical analysis of some of the more obvious forces. In this study five different factors were found to be significant: (1) the size of the corporation, (2) the extent of management control, (3) the financial connections of the corporation, (4) the relationship with competitors, and (5) the existence of local economic interests.

Size: The largest corporations tend to

TABLE 3—DISTRIBUTION OF INTERLOCKS BY KIND OF BUSINESS, 1965

Number of interlocks per corporation	Industrial	Merchandising	Transportation	Utility	Nonfinancial	Banks	Life insurance	Financial	Total
0	11	1	0	4	16	0	1	1	17
1–5	27	2	8	25	62	8	5	13	75
6–10	30	3	6	14	53	4	4	8	61
11–15	29	3	5	3	40	7	2	9	49
16–20	13	0	4	2	19	5	2	7	26
over 20	5	1	2	2	10	8	4	12	22
Total companies	115	10	25	50	200	32	18	50	250
Average number of interlocks	9.1	9.7	10.6	6.2	8.6	16.1	13.6	15.2	9.9

have the most interlocks (Table 4). This may occur because the directors of the largest corporations are the most knowledgeable, the most capable, and the most accomplished men available. Other corporations would naturally seek their advice and would rather have them on their board than men of less ability. This may also occur, however, because of factors unrelated to managerial ability. The director of a giant corporation undoubtedly has more personal influence with other companies, with potential investors, and with the government than the common man. Having the director from a large corporation on your board may also lead to profitable business with that corporation.

Management control: Management controlled companies, where management control is measured by the proportion of officers on the board of directors,[4] tend to avoid interlocks with other corporations.

[4] Management control is a qualitative matter that can not be directly measured. A. A. Berle and G. C. Means [2] and R. J. Larner [8] measured management control by the percentage of stock held by a single interest. Using 1963 data, Larner classified a corporation as being controlled by its management if no interest held 10 percent or more of its stock. By this criterion 169 of his 200 largest nonfinancial corporations were management controlled. There is, however, no significant relation between his 10 percent stockownership criterion and the officer-director criterion. Both the Berle and Means and the Larner studies may simply indicate how little stock is required to control a large corporation rather than how many large corporations are management controlled.

TABLE 4—AVERAGE NUMBER OF INTERLOCKS BY SIZE OF CORPORATION, 1965[a]

	Assets in billions of dollars							
	less than 0.5	0.5 to 0.9	1.0 to 1.4	1.5 to 1.9	2.0 to 2.9	3.0 to 3.9	4.0 to 4.9	5.0 and above
Nonfinancial	6.0	7.5	7.6	9.2	13.6	14.6	16.0	17.3
Financial	—[a]	—[a]	4.3	9.5	10.3	18.0	21.0	26.8
Total	6.0	7.5	6.8	9.2	12.4	16.4	19.1	23.7

[a] No financial corporations with assets under $1 billion are among the largest 50.

Note: The simple correlation coefficients for number of interlocks in relation to size are .316, .489, and .467 for nonfinancial, financial, and all 250 corporations, respectively. However, since the relationship does not appear to be linear, these coefficients tend to understate the degree of correlation.

TABLE 5—AVERAGE NUMBER OF INTERLOCKS IN RELATION TO MANAGEMENT CONTROL, 1965

	Percentage of directors who are officers[a]									
	less than 10	10 to 19	20 to 29	30 to 39	40 to 49	50 to 59	60 to 69	70 to 79	80 to 89	90 to 100
Nonfinancial	7.1	7.7	10.1	9.7	7.1	6.6	1.7	3.5	2.0	b
Financial	19.0	11.0	11.5	c						
Total	13.6	8.6	10.2	9.7	7.1	6.6	1.7	3.5	2.0	

[a] Officers are defined to exclude the chairman of the board.
[b] No nonfinancial corporations had 90 percent or more of their board of directors made up of officers.
[c] No financial corporations had 30 percent or more of their board of directors made up of officers.
Note: The simple correlation coefficients for number of interlocks in relation to management control are −.195, −.291, and −.292 for nonfinancial, financial, and all 250 corporations, respectively. For the 115 manufacturing corporations alone the coefficient is −.422. However, the relationship does not appear to be linear.

The frequency of interlocks with other corporations declines as the proportion of active company officers (president, vice president, treasurer, etc.) on the board of directors increases (Table 5).[5,6]

[5] These data are reported in Standard and Poor's *Register* [20].

[6] Size (as measured by assets) and management control (as measured by the percentage of directors who are officers) are virtually independent. The simple correlation coefficients for the industry subgroups are the following; manufacturing and merchandising, .090; transportation, −.032; utilities, −.110; and all financial, .002.

Financial interlocks: For nonfinancial corporations roughly one third of all interlocks are with financial institutions. For financial corporations the proportion is somewhat less (Table 6). These financial interlocks occur for several reasons.

First, companies that are in financial difficulty, particularly those occasionally threatened with insolvency, tend to form a close association with one or more financial houses. By electing a banker to the board of directors, a company may expect

TABLE 6—INTERLOCKING DIRECTORATES AMONG THE 250 LARGEST U.S. CORPORATIONS, 1965

	Number of interlocks	Interlocks with financial institutions	Interlocks with competitors[a]	Interlocks within the same commercial center
Industrial	1049	378	133	500
Merchandising	97	32	0	51
Transportation	265	90	25	63
Utilities	309	116	2	148
Total nonfinancial	1720	616	160	762
Banks	514	82	82	283
Life insurance	246	55	55	117
Total financial	760	137	137	400
All 250 corporations	2480	753	297	1162

[a] For financial institutions interlocks with other financial institutions and with competitors are identical.

to have more ready access to bank funds, while the banker can watch over the operation of the company and reduce the risk of lending to a distressed borrower.

Second, banks apparently find it advantageous to become associated with large companies by electing company officers to the bank's board of directors. This may attract large deposits as well as secure a reliable customer for bank loans.

A multiple regression analysis of the relationship between the financial position of the 200 nonfinancial corporations and the number of interlocks with financial corporations indicates that the incidence of interlocks between the two increases as the nonfinancial corporation becomes less solvent and as the assets of the nonfinancial corporation become larger.[7] Thus, many corporations are partially dependent on financial houses for credit and in turn, financial institutions depend on the larger corporations for a substantial portion of their business.

While the modern corporation typically finances a large proportion of its new investment out of internally generated funds, the volume of outside financing is still large. For all 200 nonfinancial corporations, total liabilities were 62.0 percent of equity. This percentage ranged from 102.0 for the 50 utilities to 50.8 for the 115 industrials. The importance of outside funds is further illustrated by the fact that on December 31, 1965, the total liabilities of the nonfinancial business sector in the Flow of Funds Accounts was \$461.9 billion. Of this \$276.1 billion was in the form of corporate bonds, mortgages, bank loans, and other loans, most of which was held by banks and life insurance companies. In turn, the business sector held \$20.3 billion on deposit in commercial banks [14, pp. 734–35].

Third, these financial interlocks also arise from the trust operations of banks [23]. The trust departments of the major banks are often the principal stockholders of the largest corporations, because they gather together with wealth of many individuals. Consequently, they gain repre-

[7] Two separate regression equations were calculated: one for all 200 nonfinancial corporations and one for the 50 utilities. Solvency was measured by the acid test ratio—quick assets (cash, marketable securities plus receivables) divided by current liabilities. This measure was chosen in preference to the current ratio, because the inventory element of current assets is often a cause of insolvency.

The original equation for all 200 corporations included 10 dummy variables to isolate, so far as possible, differences in the acid test which arise from differences in the kind of business. The 10 industries singled out were airlines, railroads, gas pipelines, electric power, telephone, merchandising, electrical equipment, chemicals, petroleum, and durable goods manufacturing, not elsewhere classified (n.e.c.). All other kinds of business were lumped together. Only the gas pipeline, electric power, petroleum, and durable goods manufacturing industries had significant coefficients at any point in a step-wise regression. The fitted equation is the following:

$$
\begin{aligned}
I_f = 3.11 - &\underset{(.398)}{.486R} + \underset{(.000071)}{.000446A} - \underset{(.86)}{1.66G} - \underset{(.54)}{1.33E} \\
&- \underset{(.61)}{1.96P} + \underset{(.471)}{.755M}
\end{aligned}
$$

$R^2 = .251$ $F = 9.21$, significant at the .000005 level where $N = 200$.

Where:

I_f—number of financial interlocks
R—acid test ratio
A—assets in billions of dollars
G—gas pipelines
E—electric power
P—petroleum
M—durable goods manufacturing, n.e.c.

While both the acid test and asset coefficients are of the expected sign, only the asset coefficient is significant at the 5 percent level. The lack of significance of the acid test coefficient can partly be explained by the fact that the sample included corporations conducting vastly different kinds of businesses and requiring vastly different acid test ratios. In order to eliminate this diversity and illustrate the importance of the acid test, a second regression was calculated for the 50 utilities—the largest group of corporations engaged in a similar kind of business. The fitted equation is the follow ing:

$$
I_f = 2.74 - \underset{(.95)}{1.80R} + \underset{(.000100)}{.000564A}
$$

$R^2 = .448$ $F = 19.1$, significant at the .00001 level where $N = 50$.

Both coefficients are significant at the 5 percent level using a one-tailed t test.

sentation on the board of directors of other corporations.

Competition: Nearly one in every eight interlocks involves companies which are competitors (Table 6).[8] The proportion is highest among life insurance companies, banks, and manufacturers; and lowest among merchandisers and utilities. While illegal under the Clayton Act, the law has not been effectively enforced,[9] so that the institution of interlocking directorates continues to provide a vehicle for restricting competition. Perhaps it is not often used. Perhaps it can easily be abandoned when antitrust spokesmen raise their voice. Nonetheless the framework exists today as it existed earlier in the century.

Local interest groups: The most prevalent type of interlock involves companies which have their head offices in the same commercial center (Table 6).[10] Indeed, almost half of the largest 250 corporations belong to one of 15 clearly identifiable local interest groups each of which is held together by a network of interlocking directorates. In other words, the interest groups reported by the National Resources Committee in 1935 still exist today, but in a modified form. Of the eight major groups identified by the Committee, five were associated with names of well known financial and industrial families (Morgan-First National, Rockefeller, Kuhn-Loeb, Mellon, and duPont). The remaining three groups could only be identified by their location (Chicago, Cleveland, and Boston) [24, pp. 160–163]. Today all have a local identity. Only one, the Mellon-Pittsburgh group, is clearly dominated by a single family, though the Rockefeller family occupies a position of primary importance in the New York group.

The 15 interest groups were identified by the number of times their members were interlocked together. They include corporations with head offices outside of the group city. Seven groups were classed as *tight-knit* because the corporations in those groups interlocked four or more times, while eight groups were classed as

[8] The FTC compiled a list for the House Antitrust Subcommittee of industrial corporations which had interlocking directors and which did business in the same five digit SIC classification [22]. While this is not conclusive evidence that they are competitors—they could do business in separate localities, for example—it is a first approximation. Moody's Manuals were used to determine whether merchandising, transportation, and utility companies were competitors. Merchandising firms were considered to be in competition if they sold the same line of goods in one common city. None of those that interlocked were competitors. Transportation firms were considered to be competitors if they served two terminal points in common. By law utilities can not compete. However, two interlocking gas pipelines were found that shipped from one common region to another. All banks and life insurance companies were assumed to be competitors, because they deal extensively in federal obligations and because most of them deal in related state, local, and corporate loans and securities. It would not be surprising, for example, to find a single large manufacturer with a line of credit at most of the 32 banks in the survey. Thus, in the case of financial institutions an interlock with a *financial interest* can not be distinguished from an interlock with a competitor.

[9] In 1965 the Antitrust Subcommittee of the House Judiciary Committee concluded after a lengthy study of interlocks among competitors that

> In operation, the body of Federal legislation has not effectively prevented interlocks in corporate managements in the fields it covers. Enforcement of the Clayton Act's prohibitions against interlocking directorates was neither prompt nor vigorous . . .
>
> From its enactment on October 15, 1914, to January 1965, the FTC had filed a total of 13 complaints under section 8 of the Clayton Act. Only one of these complaints resulted in a cease-and-desist order, and this was by consent; the remainder were dismissed when the directors involved discontinued the prohibited relationship.
>
> The Department of Justice did not undertake a systematic program with respect to interlocking directorates until after World War II, and the first cases to be litigated to a decision by a court were not filed until February 27, 1952, 38 years after the enactment of the Clayton Act. As of January 1965, the Department of Justice had instituted a total of 10 cases to enforce section 8, and 5 cases to enforce section 10 [22, pp. 226–227].

[10] A commercial center was defined to include a metropolitan area and its immediate environs so that suburban head offices would not be left out.

TABLE 7—INTEREST GROUPINGS

Tight-knit groups:[a]	Number of Corporations
New York	38
Chicago	14
San Francisco	13
Pittsburgh	8
Los Angeles	6
Cleveland	5
Detroit	4
Loose-knit groups:[b]	
Hartford	6
Philadelphia	4
Milwaukee	7
Portland, Ore.	4
Minneapolis-St. Paul	3
Boston	3
Dallas	2
Houston	2
Unallocated[c]	3
Total	122

[a] Interlocked four or more times, except for New York where all corporations were interlocked six or more times.
[b] Interlocked two or three times.
[c] Interlocked four or more times with two groups and not allocated to either group. The three are Pan American World Airways, the Pennsylvania Railroad, and Union Oil.

loose-knit because they interlocked only two or three times. New York is actually in a class by itself. So many corporations interlocked with the New York group that it was necessary to raise the cut-off point for membership in the group to six or more interlocks (Table 7).

Nearly all the groups share certain common characteristics. Banks or life insurance companies form the central core of the group and have the greatest number of interlocks with other members of the group. Local public utilities form a second ring about the central core and have the second greatest number of interlocks. Finally, an outer ring is made up of manufacturing, merchandising, and transportation companies that do a substantial portion of their business in the region of the group city.

The New York group is, in part, an exception to this general pattern. It contains the major New York banks, life insurance companies, and utilities, but it also contains a large number of companies whose business is clearly nationwide.[11]

In 1935, the National Resources Committee designated two major New York City groups, the Morgan-First National and the Rockefeller. Today it is not possible to separate these groups. For that matter, it has not been easy to distinguish between them since the turn of the century, as John Moody observed in 1904:

> It should not be supposed, however, that these two great groups of capitalists and financiers are in any real sense rivals or competitors for power, or that such a thing as a "war" exists between them. For, as a matter of fact, they are not only friendly, but they are allied to each other by many close ties, and it would probably require only a little stretch of the imagination to describe them as a single great Morgan-Rockefeller group [10, pp. 492–93].

The Chicago group and the Pittsburgh group contain many of the companies they did in 1935. The Chicago group is made up of 14 corporations that have been prom-

[11] The New York group centers around six large banks—[the number in the parentheses indicates the number of times the corporation interlocks with others in the same group] Chase Manhattan (22), First National City (14), Chemical Bank (18), Manufacturers Hanover Trust (20), Bankers Trust (10), and Morgan Guaranty Trust (11); four life insurance companies—Metropolitan (19), Equitable Life Assurance (23), New York Life (15), and Mutual of New York (15); and several long established industrials and utilities—A.T.&T. (13), Consolidated Edison (15), U. S. Steel (11), General Electric (12), Union Carbide (12), General Foods (13), International Paper (11), Phelps Dodge (11), Corn Products (10), Chrysler (15), American Smelting and Refining (11), U. S. Rubber (10), Ford Motor (10), and National Distillers and Chemical (10). Each of these 24 corporations interlock 10 or more times with other corporations in the group. The following corporations complete the group: National Dairy Products (8), Allied Chemical (6), IBM (9), B. F. Goodrich (7), Irving Trust (7), American Electric Power (6), Southern Railway (9), Union Pacific (9), Panhandle Eastern Pipe Lines (6), Socony Mobil Oil (6), Western Electric (7), F. W. Woolworth (6), Borg-Warner (7), and Texaco (9).

inent in the economic history of the city,[12] while the Pittsburgh group includes the principal corporations that have long been associated with the Mellon name.[13] San Francisco did not appear as a group in the 1935 study, however, today it is third in size. Like Chicago, the San Francisco group is dominated by local banks and utilities and by other important regional corporations.[14] The remaining groups follow the same general pattern.[15]

The arbitrary rules used to establish these groups involves three serious problems. First, 15 corporations qualified for membership in two groups, none qualified for membership in three. Such inter-group corporations were not allocated to any group unless there was a clear connection to a particular group in terms of location, ownership, or number of interlocks.[16] Second, some corporations were included in groups with which they have little in common (location, products, ownership) and from which they may be independent in every respect except the coincidence of their common directors. General Motors, for example, is included in the Pittsburgh group simply because two directors sit on its board who are multiple directors in the Pittsburgh group. Third, many significant interlocks and interlocking groups were omitted by restricting the study to just the 250 largest corporations. W. L. Warner, D. B. Unwalla, and J. H. Trimm have found in a sociological study of interlocks that 5776 directors of their 500 large representative corporations interlocked a total of 8872 corporations, ranging in size from less than a million dollars in net worth to over a billion dollars in assets [13, pp. 130–34].

[12] The Chicago group includes three major banks—Continental Illinois (17), First National Bank of Chicago (13), Harris Trust (9); two utilities—Commonwealth Edison (11) and Peoples Gas Light and Coke (5); two meat packers—Swift (14) and Armour (6); two mail-order/department store chains—Sears, Roebuck (10) and Montgomery Ward (6); plus International Harvester (21), Standard Oil of Indiana (6), Pure Oil (4), Inland Steel (10), and the Illinois Central (11).

[13] The Pittsburgh group contains: Mellon National Bank (21), Gulf Oil (13), Alcoa (8), Westinghouse Electric (7), Consolidation Coal (5), Jones and Laughlin (9), and Pittsburgh Plate Glass (5). The group also includes General Motors (6).

[14] The San Francisco group is made up of the following: Bank of America (5), Wells-Fargo (7), Crocker-Citizens (7), Bank of California (8), Pacific Gas and Electric (11), Pacific Lighting (4), Kaiser Aluminum and Chemical (7), Kaiser Industries (4), Safeway Stores (5), FMC (6), Southern Pacific (10), Tennessee Gas Transmission (7), and Caterpillar Tractor (7).

[15] The remaining groups are the following: Los Angeles, United California Bank (10), Security First National (4), Southern California Edison (12), North American Aviation (7), Lockheed Aircraft (5), and American Metal Climax (4); Cleveland, Cleveland Trust (6), Cleveland Electric Illuminating (6), Republic Steel (4), Baltimore and Ohio (6), Chesapeake and Ohio (4); Detroit, National Bank of Detroit (4), Detroit Bank and Trust (4), Detroit Edison (4), and National Steel (5); Hartford, Travelers (4), Aetna (3), Connecticut General (2), Connecticut Mutual (2), Southern New England Telephone (8), and United Aircraft (9); Philadelphia, First Pennsylvania Banking and Trust (4), Philadelphia National Bank (2), Penn Mutual (6), and Philadelphia Electric (3); Milwaukee, Northwestern Mutual (9), Wisconsin Electric Power (3), Allis-Chalmers (8), Kimberly-Clark (5), American Can (2), American Natural Gas (3), and Chicago, Milwaukee, St. Paul and Pacific (2); Portland, Oregon, U. S. National Bank of Oregon (4), Pacific Power and Light (2), Weyerhaeuser (3), and the Great Northern (3); Minneapolis-St. Paul, Northern States Power (2), Chicago, Burlington and Quincy (2), and the Northern Pacific (2); Boston, First National Bank of Boston (4), John Hancock Mutual (2), and New England Life (2); Dallas, Republic National Bank of Dallas (2) and Texas Utilities (2); and Houston, El Paso Natural Gas (2) and Transcontinental Gas Pipe Lines (2).

[16] Thirteen of the 15 inter-group corporations were interlocked four or more times with the New York group. They were the following [the first number in parentheses indicates the number of New York interlocks, the second number indicates the number of interlocks with the second group]: New York-Chicago, Chase Manhattan (22,4), New York Life (15,4), and First National Bank of Chicago (6,13); New York-San Francisco, Southern Pacific (10,10); New York-Pittsburgh, Mellon National Bank (4,22), Westinghouse Electric (6,7), Consolidation Coal (7,5) and Pan American World Airways (6,5); New York-Detroit, Chrysler (15,4); New York-Hartford, Chase Manhattan (22,4), First National City (14,4), Travelers (10,4), and United Aircraft (6,9). The remaining two inter-group corporations were the following: San Francisco-Los Angeles, Union Oil (5,6); and Pittsburgh-Philadelphia, Pennsylvania Railroad (5,6).

While these three definitional problems affect the number, size and membership of groups, it is doubtful that they affect the nature of interlocking groups in general. The *tight-knit* groups in particular are intertwined too many times to be significantly rearranged by minor changes in definition.

III. *Conclusion*

The institution of the interlocking directorate is extensive and enduring. Most of the larger corporations have been interlocked with other large corporations for many decades. This suggests that the structure of the American economy is markedly different from what is commonly supposed. The widely accepted views of A. A. Berle and G. C. Means [2], R. A. Gordon [7], R. J. Larner [8], and J. K. Galbraith [6] that the modern corporation is an independent and self-sufficient organ ruled by its own self-perpetuating management needs to be modified on several points.

The extreme view holds that: "Major corporations in most instances do not seek capital. They form it themselves [3, p. 40]." This view contains an important element of truth, but it overlooks the fact that the total liabilities of nonfinancial business approach one-half trillion dollars, that about one-third of the assets of the 200 largest nonfinancial corporations are financed on credit, and that these 200 corporations interlock 616 times with the 50 largest banks and life insurance companies alone. Stock and bond issues, mergers and acquisitions, and other questions of high finance require expert counsel. Such questions are not the daily business of the salaried executives of nonfinancial corporations, the men who Gordon claims make "... the essential business decisions ... [7, p. viii]," nor do the anonymous men of Galbraith's "technostructure" have the opportunity to develop competence in handling such occasional and specialized questions. This does not mean that a small clique of bankers controls every detail of corporate activity. However, the presence of knowledgeable men of finance on the board of directors can not help but influence policies within the sphere of their competence and responsibility.

The presence of outside local business leaders on the board of directors must also force management to consider the interests of the local community, both in terms of its economic growth and in terms of its social and political development. In addition, the presence of competitors in the board room must direct the attention of the management to certain matters of common interest.

Thus, while it is accepted that the modern corporation is the central unit of production and capital accumulation today, its autonomy is a matter of degree. Its autonomy increases as management control over the board of directors increases, for then management can isolate itself from other points of view. For the typical corporation this control is far from absolute. Within its own walls it faces the constraining influence of the financier, the local interest, and the competitor.

References

1. P. A. Baran and P. M. Sweezy, *Monopoly Capital.* New York 1966.
2. A. A. Berle and G. C. Means, *The Modern Corporation and Private Property.* New York 1932.
3. A. A. Berle, *The 20th Century Capitalist Revolution.* New York 1954.
4. L. D. Brandeis, *Other People's Money.* New York 1914.
5. G. W. Domhoff, *Who Rules America?* Engelwood Cliffs 1967.
6. J. K. Galbraith, *The New Industrial State.* Boston 1967.

7. R. A. GORDON, *Business Leadership in the Large Corporation.* Berkeley 1945.
8. R. J. LARNER, "Ownership and Control of the 200 Largest Nonfinancial Corporations, 1929 and 1963," *Am. Econ. Rev.*, Sept. 1966, *56*, 777–787.
9. C. W. MILLS, *The Power Elite.* New York 1959.
10. J. MOODY, *The Truth about the Trusts.* New York 1904.
11. V. PERLO, *The Empire of High Finance.* New York 1957.
12. D. VILLAREJO, "Stock Ownership and Control of Corporations," *New University Thought*, Autumn 1961, 33–77.
13. W. L. WARNER, D. B. UNWALLA and J. H. TRIMM, *The Emergent American Society.* New Haven 1967.
14. Board of Governors of the Federal Reserve System, *Federal Reserve Bulletin*, May 1966, *52*, 724–735.
15. FORTUNE, *The Fortune Directory: The 500 Largest Corporations.* New York, 1965.
16. MOODY'S INVESTORS SERVICE, *Moody's Bank and Finance Manual, 1965.* New York 1965.
17. ——————, *Moody's Industrial Manual, 1965.* New York 1965.
18. ——————, *Moody's Public Utility Manual, 1965.* New York 1965.
19. ——————, *Moody's Transportation Manual, 1965.* New York 1965.
20. STANDARD and POOR'S, *Poor's Register of Corporation Officers and Directors, United States and Canada, 1965.* New York 1965.
21. U. S. FEDERAL TRADE COMMISSION, *Report on Interlocking Directorates.* Washington 1951.
22. U.S. HOUSE OF REPRESENTATIVES, COMMITTEE on the JUDICIARY, Antitrust Subcommittee, *Interlocks in Corporate Management.* Washington 1965.
23. ——————, COMMITTEE on BANKING and CURRENCY, Subcommittee on Domestic Finance, *Commercial Banks and Their Trust Activities: Emerging Influence on the American Economy.* Washington 1968.
24. U. S. NATIONAL RESOURCES COMMITTEE, *The Structure of the American Economy: Part I. Basic Characteristics.* Washington 1939.

The Optimum Lifetime Distribution of Consumption Expenditures

By LESTER C. THUROW*

The actual lifetime distribution of consumption expenditures may have little relation to the optimum lifetime distribution. Individuals can easily redistribute consumption into the future by saving, but they cannot easily borrow for present consumption. The high risks associated with expected earnings make borrowing difficult. Without slavery there is no effective way to mortgage expected future earnings. Consequently the current flow of income rather than the total lifetime flow of income may dominate current consumption expenditures. The individual is unable to optimize his distribution of consumption expenditures.

The importance of the institutional constraints upon lifetime redistribution depends upon how much individuals wish to redistribute consumption expenditures away from their actual income stream. For the average person consumption and expenditures move closely together (see Figure 1), but this may not be by choice.[1] Since there is no a priori method to determine the optimum distribution of consumption, this paper develops an empirical technique for estimating the optimum lifetime distribution of consumption expenditures.

I. *The Theory*

Given an expected lifetime budget constraint, an individual maximizing utility will arrange his lifetime consumption pattern in accordance with the well known Fisherian [1] conditions (see Equation 1).[2]

$$\text{(1)} \qquad \frac{MU_c^i}{MU_c^j} = (1 + r_1 - r_2)^{j-i}$$

where:

MU_c^i = marginal utility of consumption in year i

MU_c^j = marginal utility of consumption in year j

r_1 = market rate of interest between years i and j

r_2 = individual rate of time preference between years i and j

The ratios of the marginal utilities of consumption in any two years must be equal to the individual's personal discount differential, $(1+r_1-r_2)^{j-i}$.[3] The personal discount differential is composed of two elements, the market rate of interest (r_1) and the individual's subjective rate of time preference (r_2). If the personal discount differential is one ($r_1-r_2=0$), the utility maximizer will arrange his consumption pattern to make the marginal utility of consumption equal in each year. Such procedures maximize expected life-

* The author is an associate professor of economics and management at the Massachusetts Institute of Technology. The ideas for this paper were developed while working on a project at the NBER. The author would like to thank the anonymous referee for his excellent comments. The remaining errors are those of the author.

[1] All data in this paper come from the Bureau of Labor Statistics [3, pp. 23–29].

[2] Fisherian conditions assume that the individual is neutral with respect to risk and that he can estimate his expected lifetime income.

[3] In the standard expression of the Fisherian conditions, the subjective rate of time preference affects the marginal utility of income schedules rather than appearing explicitly. Consequently the marginal utility of income schedule for any year depends on the year from which it is being viewed. This is not the case under the formulation in Equation (1). The marginal utility of income schedule in any year is invariant to the year from which it is being viewed. The influence of time preference appears explicitly.

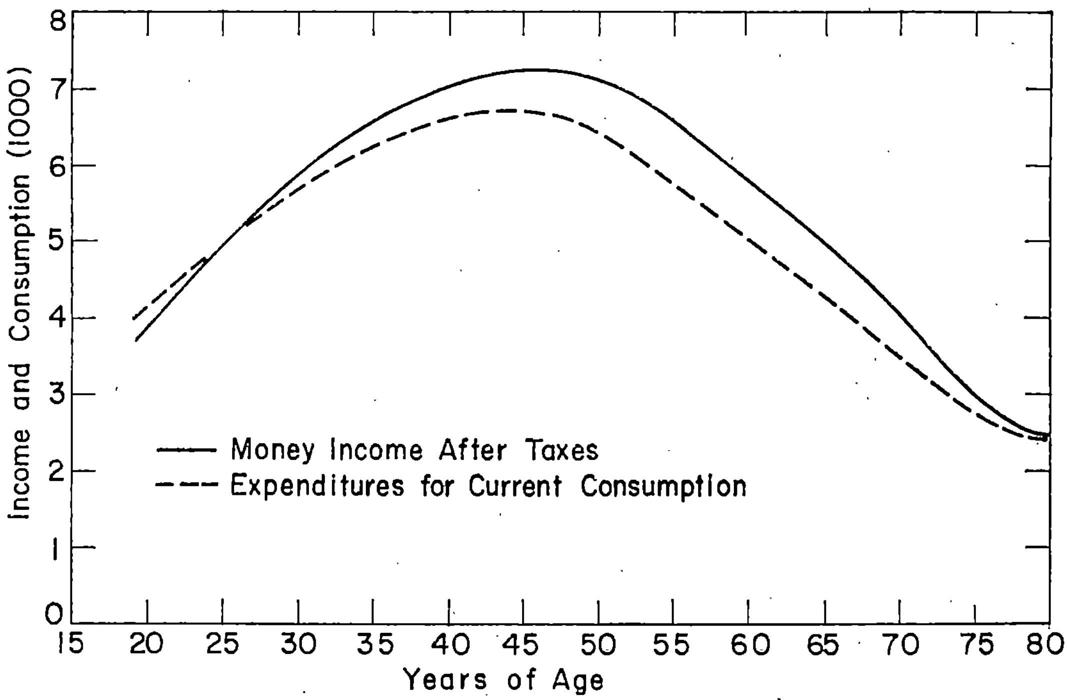


FIGURE 1. AVERAGE MONEY INCOME AFTER TAXES AND EXPENDITURES FOR CURRENT CONSUMPTION.

time utility, but they do not necessarily lead to equal consumption in all periods.

An individual may have a varying positive or negative rate of time preference. If he prefers to consume goods and services while young, his rate of time preference is positive. Enjoyment may also fluctuate over a lifetime rather than being a smooth function of time, but the higher the rate of time preference and the lower the market rate of interest between any two years, the more consumption will occur in the earlier year.

Consumption will also be higher in those years where the marginal utility of consumption expenditures are higher. Needs differ over the course of a lifetime.[4] Consumption of medical services rises at particular times of life; expenditures on children occur while an individual is raising his family. Education expenditures may be high when young and again when children are being educated. Other expenditures are lumpy. Durable goods purchases are large when setting up a household. If these goods are to be enjoyed, they must be purchased in a limited period of time.

If individuals save only for future consumption,[5] positive savings indicate that the individual wishes to transfer some of his income from present consumption to future consumption. Negative savings indicate that the individual wishes to transfer consumption from the future to the present. If a person is neither dissaving or saving, he has reached his desired consumption level given his personal discount differential and his consumption preferences.

As mentioned there are institutional constraints on dissaving. They are not

[4] A rising probability of death may reduce the utility of consumption while elderly, if the individual is a risk averter.

[5] In this paper, future consumption is assumed to be the only motive for saving.

absolute, however. Dissaving is severely limited but not prevented.[6] Consequently empirical observations on dissaving cannot indicate how much the individual would like to dissave, but they do indicate that the individual is not at his desired consumption level. He wishes to consume more than his current income allows. Individuals will continue to dissave as their income rises until current income equals desired current consumption. Thus if it were possible to estimate the income at which an individual becomes a zero saver for each year of his life, it would be possible to determine his optimum lifetime distribution of consumption.

Individuals undoubtedly differ, but consumer expenditure and income surveys make it possible to determine the income at which the average person in a particular age group neither saves nor dissaves. Using these surveys it is possible to determine the average desired lifetime distribution of consumption expenditures. The technique, however, has several limitations or problems.

(1) Using cross section data, the actual average lifetime income does not equal the lifetime income calculated by summing the income levels at which savings are zero. If the lifetime distribution of income is independent of the absolute level of lifetime income, the problem can be solved by scaling consumption expenditures at each age up or down by a constant fraction so that they sum to the actual average lifetime income. With this assumption the marginal utility of consumption may still depend upon total consumption, but the ratios of consumption at different points in a lifetime are assumed to be independent of lifetime income. The proportion of total consumption spent in any one year is assumed to remain constant as total consumption varies. This may be untrue. A higher lifetime income, for example, may lead to a more than proportionate increase in education expenditures (and total expenditures) while young. For lack of a better assumption, a constant scaling fraction is used in the following empirical analysis. Some empirical evidence for this assumption is presented in footnote 14.

(2) Using cross section data assumes that individuals in older age brackets are similar in all relevant aspects with those in younger age brackets. This is obviously not true. The average older individual has a lower lifetime income than the average young person. Ignoring this factor biases the results toward too much consumption in the early years of an individual's lifetime since it is assumed that the young will consume the same amounts when elderly as the current elderly. In all probability they will consume more since they have a higher lifetime income. There may be other systematic differences which affect consumption patterns. Education may differ, time preferences may differ, etc. These factors bias the results, but the direction of the bias cannot be determined theoretically.

(3) If total dissaving as well as the annual increments to dissaving are limited, savings may occur to repay previous debts even though the individual would like to continue dissaving. Thus a position of zero savings would not indicate a position of equilibrium. This would bias the results towards too little consumption while young. Empirically this possibility does not seem to be an acute problem. Individuals are increasing assets rather than reducing liabilities at the point where they become zero savers.

(4) Individuals within the same age group are not homogeneous. Different individuals may have very different patterns of income. Laborers reach an income plateau much earlier than professional workers. Thus it may be possible to have

[6] In every age group there is substantial amounts of dissaving at low income levels.

an income class composed of laborers with balanced budgets while higher income classes are dissaving because they are composed of professional workers. This would bias the results toward too little consumption while young. Empirically there are no age groups with double points of equilibrium or where savings decline as income increases, but there may be some biases in this direction.

(5) The income class with zero saving in aggregate may be composed of individuals who are either large savers or dissavers. They simply cancel each other. The dissaving of professional workers with higher lifetime incomes and more slowly rising incomes could exactly balance the saving of laborers with lower lifetime incomes and faster rising incomes. In this case the average optimum lifetime distribution of consumption might not look like the optimum distribution of consumption for any actual person. It is simply a weighted average of two distinct curves where the weights depend not only on the numbers of individuals but their relative incomes. If laboring income is higher relative to professional income while young, the results will be biased toward too little consumption while young and too much consumption while elderly.[7]

(6) The procedure assumes that the lifetime utility function has annual utility levels as its sole arguments. In addition the annual utilities are assumed to be independent and additive. Consumption in year one is assumed not to affect the utility of consumption in year two. If there are complementarities between consumption at different points in time, it is impossible to compare individuals at different points of time unless they have the same history of previous and future consumption.

[7] An alternative, if the sample were larger, is to find the central tendency of all those families with balanced budgets.

(7) Since goods need not be totally consumed in the year in which they are purchased, the procedure estimates the optimum distribution of consumption expenditures and not the optimum distribution of consumption.

(8) Given optimum consumption expenditures in any two years, it is not possible to separate the effects of the personal discount differential and differences in the marginal utilities of income. There are many combinations of marginal utilities and personal discount differentials that are compatible with the estimated equilibrium.

II. *The Data*

The Bureau of Labor Statistics' *Survey of Consumer Expenditures and Income in 1960–61* [3, pp. 23–29] provides the data base. Since there are errors in the BLS survey any systematic over- or under-reporting for different age groups of either income or consumption expenditures will bias the results.

Two definitions of zero savings are used.[8] In the first, savings are zero when expenditures for current consumption equal money income after taxes. In the second, savings are zero when expenditures for current consumption plus gifts and contributions plus personal insurance payments equal money income after taxes.

The second definition was used for two reasons. (1) Positive consumption benefits are derived from being able to provide gifts and contributions to family members, friends, and charities. Benefits may also vary over the lifetime of an individual. Consequently gifts might be considered consumption expenditures. (2) No allowance has yet been made for risk and uncertainty. If risk and uncertainty differ

[8] Linear interpolation was used to find the precise point of equilibrium and free hand interpolation was used to interpolate optimum consumption curves between the midpoints of income classes.

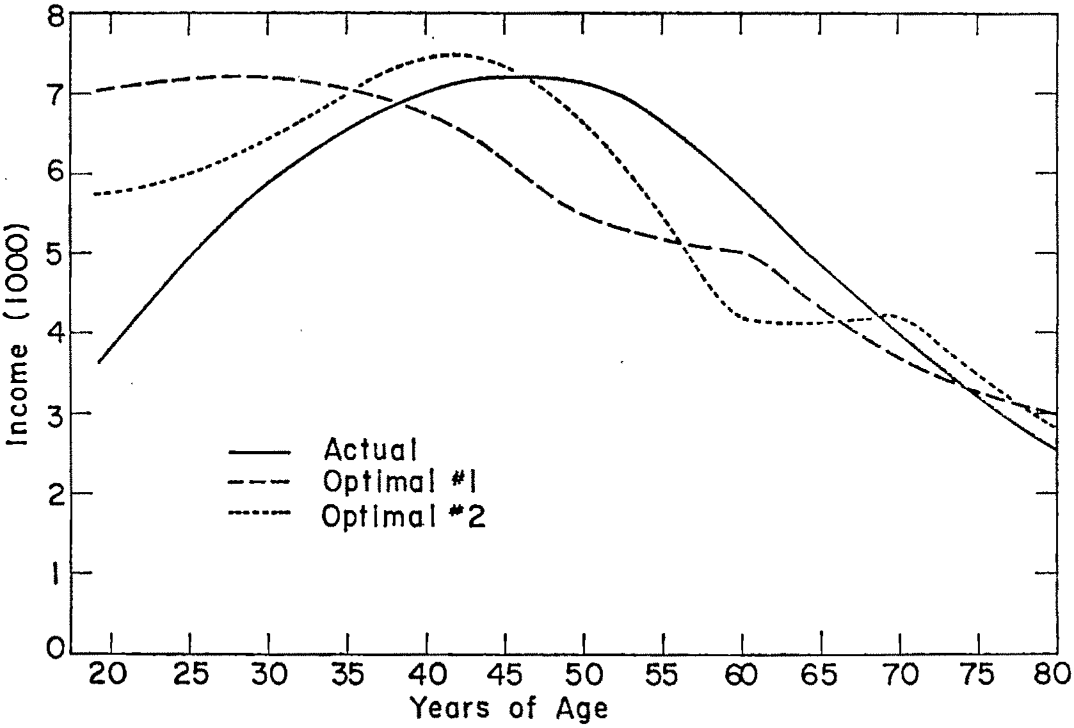


FIGURE 2. ACTUAL AND OPTIMAL DISTRIBUTIONS OF INCOME FOR URBAN FAMILIES.

systematically by age, the marginal utilities of income might not be equal at the point where consumption equals income. At different ages the individual must be making provisions for different amounts of risk. Risk may lead to a need for liquid assets and the consequent depression of consumption expenditures. There is no a priori method to determine when risk is highest. Risks may be higher when old since the probability of death and sickness is higher, but they might equally well be higher when young since fewer assets have been accumulated and there are more dependents who would suffer for a longer period of time.

To the extent that insurance premiums are not savings but are payments for pure insurance designed to eliminate the effects of risk and uncertainty, insurance premiums are consumption expenditures to reduce risk and uncertainty. If the insurance market is in equilibrium, including insurance premiums in consumption expenditures should correct for differences in risk at different ages.[9]

III. *The Results*

Using the first definition of zero saving, a desired distribution of consumption emerges which is quite different from the actual distribution of income (see optimum distribution #1, Figure 2).[10] If the family head lives from age 19 to 80, the average family has a lifetime income of $342,246 in 1960–61. The average family prefers to redistribute $30,242 into the years be-

[9] If the proportions of observed insurance premiums that are pure insurance and the proportion that are savings vary over a lifetime, this will bias the results toward too much consumption in years with a greater than average amount of insurance savings.

[10] Since potential dissavings is small and all income is not actually consumed the optimum distributions of consumption will be compared with the actual distribution of income. The distribution of income is a good estimate of potential consumption possibilities.

tween 19 and 35 and $1560 into the years between 75 and 80 by borrowing from the years between 35 and 75.[11] Instead of having its highest income years between 40 and 55 years of age, the average family prefers to have its highest income years between 20 and 35 years of age. Possible reasons are not hard to find. During the 20 to 35 age period, the family is faced with problems of raising a family and building up its stock of consumers durables. Enjoyment of consumption goods may also be highest in these years.

The second definition of zero savings provides a slightly different optimum consumption curve (see optimum distribution #2). The average family prefers to redistribute $18,134 into the years between 19 and 46, and $3484 into the years between 69 and 80. The total amount of redistribution falls from $31,802 under the first set of assumptions to $21,618 under the second. Less is desired in the early years; more is desired in the later years. Allowing for risks, uncertainty, and gifts leads to relatively less need for income between 19 and 35 and relatively more need between 35 and 46 and between 69 and 75.[12]

IV. *The Conclusions*

Both sets of assumptions yield the same basic conclusions. Families desire a substantial amount of lifetime income redistribution over and above that done in 1960–61 and this redistribution is heavily weighted toward the younger years of a family's life.[13,14] The actual lifetime pattern of income is a severe constraint on the desired lifetime distribution of consumption expenditures. Based on the results of this analysis, lifetime welfare levels might be substantially increased if the constraints on lifetime income redistribution could be lifted. Consequently social planners should investigate methods of eliminating the institutional constraints.

Perhaps proposals like the educational opportunity bank [2] ought to be expanded to permit loans for current consumption expenditures other than education.[15] Consumption loans could be secured through incurring a liability to pay a higher income tax rate until repayment is made. Only a governmental institution such as the educational opportunity bank can provide a method of securing expected future earnings as collateral and be large enough to eliminate the risks associated with the large variance in the income distribution of any particular group. As a second best solution, the Social Security

[11] In terms of time preference, there is a positive rate of time preference between the ages 19–35 and the ages 35–75, and a negative rate of time preference between the ages 35–75 and the ages 75–80. These rates of time preference, however, are not based on a zero rate of interest. They reflect the rates of interest actually existing in the economy in 1960–61.

[12] Since there is a substantial amount of insurance savings and since the proportion probably rises with age, these results are probably much closer together than the numbers would indicate.

[13] This conclusion might be modified if there were systematic differences in inherited wealth by age group. Greater wealth would lead to higher consumption. Although wealth is a possible source of bias, it should not be important. The survey contains no information on wealth, but it was supposedly drawn to avoid such problems.

[14] The two definitions of zero saving make it possible to provide some information to substantiate the assumption that the ratios of consumption expenditures are independent of total lifetime consumption. With the first definition of consumption, a lifetime income of $254,946 is necessary to insure that money income after taxes equals expenditures for current consumption at each point in time. With the second definition of consumption an income of $423,621 is necessary. Since the actual average lifetime income in the survey was $342,246, annual consumption expenditures are scaled up under the first definition and down under the second. Since both definitions indicate that families desire substantial amounts of lifetime income redistribution and that this redistribution is heavily weighted toward the younger years, these qualitative results are invariant with respect to whether the results are scaled up or down. Thus the central conclusions only depend on the assumption that the distribution of consumption changes monotonically between income levels of $254,946 and $423,621. Within this constraint ratios may vary without altering the qualitative conclusions.

[15] For a detailed description of the educational opportunity bank, see Schell [2].

An Axiomatic Model of Logrolling

By ROBERT WILSON*

Logrolling, or the trading among legislators of their votes on the different bills before an assembly, has been cited in several recent studies as a redeeming feature of democratic processes for social choice; mainly because it allows expression of intensities of preferences by the legislators. For example, a legislator might offer to another to trade his vote on a certain bill in return for the other's votes on a second and a third bill. The prominence of the logrolling hypothesis in political science seems to derive from the early exposition of Arthur Bentley [2], and more recently economic analyses have been provided by James Buchanan and Gordon Tullock [4], James Coleman [5] [6], Dennis Mueller [9], and R. E. Park [10].

However, in commenting on the arguments of Buchanan and Tullock, Kenneth Arrow [1, p. 109] has noted that the possibility of logrolling in no way invalidates the conclusion of his *General Possibility Theorem.* Arrow's characterization of logrolling, nevertheless, depends upon a presumption which may or may not obtain in any specific instance. Arrow notes that if the bills before an assembly, or the legislators' preferences for the bills, are mutually dependent then the voting process for the bills is equivalent to ballots between alternative social states, and hence [1, p. 48], it is only the transitivity requirement of a social welfare function that is violated.

In this paper we want to consider the alternative case in which each voter's preferences for the bills are independent, in order to investigate the possibility of a different characterization of logrolling in relation to the *General Possibility Theorem.* In fact, for a special case of an axiomatic model which is similar to Arrow's, we shall find that logrolling violates Arrow's Condition 3, the Independence of Irrelevant Alternatives. This result is of some interest, in part because Arrow [1, p. 27] has remarked that the Independence of Irrelevant Alternatives " . . . implies that in a generalized sense all methods of social choice are of the type of voting . . . , " and in part because this condition has most often been doubted by Arrow's critics. We obtain an exact characterization of the way in which this condition is violated, and an example is provided.

Our model also affords a feature not previously encountered in social choice theory; viz., the conditions imposed on the social choice imply more severe restrictions on the individual voters' preferences than were initially assumed. If this curiosity were general it would raise new questions about the conditions that can acceptably be imposed on social choice processes. In the present case, this result derives from a theorem demonstrating that (in a special case) each voter's allocation of his voting strength among the bills must constitute an additive utility representation of his preferences; i.e., his voting must be sincere. The proof depends upon

* The author is a member of the faculty of the Graduate School of Business, Stanford University. This study was prepared during the tenure of a Ford Foundation Faculty Research Fellowship and with the support of the Center for Operations Research and Econometrics at the Catholic University of Louvain, Belgium, where in an earlier version it appeared as a working paper, "Arrow's Possibility Theorem for Vote Trading," which was presented at the Conference on the Mathematical Theory of Committees and Elections organized by the Institute for Advanced Studies, Vienna, June 26–28, 1968.

individualistic action and the assumed unimportance of the status quo.

I. *An Exchange Economy for Vote Trading*

The context of our study is an assembly in its simplest setting, devoid of the complications of parliamentary procedure, internal committees, an upper house, a presidential veto, and the dynamics of new bills and amendments; hence, we shall call it a referendum. The referendum consists of a fixed collection of bills proposed for passage, say n in number, comprising the set $J=\{j \mid 1 \leq j \leq n\}$. Several citizens, comprising the set $I=\{i \mid 1 \leq i \leq m\}$, are to vote on the bills, and each citizen is initially endowed with one vote on each bill. A bill will be passed if the number of affirmative votes it receives exceeds the number of negative votes, each citizen being free to declare whether a vote he controls on a bill is to be cast as affirmative, neutral, or negative. Moreover, to allow for the possibility of logrolling, or vote trading, we suppose that each citizen can transfer his control of a vote on a bill to another citizen as he chooses; however, no side payments (for example, in money) are possible or permitted.

The process of vote trading in the situation outlined above is essentially a barter economy in which the citizens' initial endowments of votes on the various bills correspond to endowments of apples, oranges, pears, etc., in the usual paradigm leading to the familiar Edgeworth box diagram of exchange possibilities. In this case, the citizens' preferences for votes of each kind derive from their preferences for the bills whose disposition they may affect. Among the many exchange possibilities on the Edgeworth contract curve that may be attained from the initial endowments, there is a subset of special interest that will provide the focus for the present analysis: these are the exchanges that may result from an exchange economy, in which all trades are made at stable equilibrium prices or exchange ratios among votes on the various bills. In an exchange economy, no trades are made at discretionary prices other than the equilibrium prices. One might, for example, suppose that votes are freely transferable commodities (see Mueller [9]) and that prices or exchange ratios are advertised publicly, with trades being accomplished anonymously through brokers, much in the fashion of a stock market.

Let Π_j denote the equilibrium price of a vote on the j-th bill, so that Π_k/Π_j is the exchange ratio between votes on the j-th and k-th bills, and adopt the normalization $\sum_{j\epsilon J}\Pi_j=1$ without any loss of generality. We can then suppose equivalently that the i-th citizen, since he is endowed with one vote on each bill, is endowed (in whatever unit of currency the equilibrium prices are measured) with a budget of $\sum_{j\epsilon J}1\cdot\Pi_j=1$ valued at the equilibrium prices, and that he can feasibly acquire any distribution of votes $\{V_{ij} \mid V_{ij}\geq 0, j=1, \cdots, n\}$ for which the budget condition $\sum_{j\epsilon J}\Pi_j V_{ij}\leq 1$ is satisfied. In total, of course, there is the limit on the supply of votes on each bill, $\sum_{i\epsilon I}V_{ij}\leq m$, which in the market determines the equilibrium prices.

In the following we shall actually employ a more compact notational description of the exchange economy for votes. Let σ_{ij} be $+1$, 0, or -1 accordingly as the i-th citizen declares his votes on the j-th bill to be cast as affirmative, neutral, or negative, and define $d_j^i=\sigma_{ij}\Pi_j V_{ij}$ and the vector $d^i=(d_j^i)$. The latter we shall call the citizen's decision rule, or voting strategy, for allocating his budget of voting power. In general, a decision rule will depend upon the citizen's preferences for the bills as well as possibly the preferences of other citizens, and various other factors that may be envisioned: when only the former are of interest we shall use the

notation $d^i(R^1, \cdots, R^m)$ to express this dependence.

Definition: A decision rule for the i-th citizen is an n-dimensional real vector d^i such that $\sum_{j\in J}|d_j^i| \leq 1$.

Note that the interpretation above of the citizens' decision rules is not confining: several citizens might form a coalition which would determine entirely or in part their decision rules.

Parenthetical Remarks about the Formulation

It should be observed, however, that in adopting a formulation in terms of the citizens' decision rules we have made a subtle change in the game-theoretic structure of the problem.[1] In the *exchange-economy* game, each citizen decides on the number of votes he will acquire on each bill, subject to a budget constraint which depends on the equilibrium prices. In the *decision-rule* game (which is essentially a *Blotto* game), each citizen decides on the allocation of his fixed budget among the bills. The latter case is equivalent to one in which each citizen is endowed with a budget of fungible ballots which can be cast on any bill. The two games are equivalent in the present instance only because: (1) since all citizens are equally endowed with votes on all bills in the *exchange-economy* game, the valuation of their endowments at equilibrium prices yields the same budgets (viz., 1 for every citizen) for all (normalized) sets of equilibrium prices; and (2) the disposition of any bill j is determined by the condition $\sum_{i\in I}\sigma_{ij}V_{ij}>0$ for passage which is exactly equivalent to the condition $\sum_{i\in I}\Pi_j\sigma_{ij}V_{ij}>0$, or $\sum_{i\in I}d_j^i>0$ in terms of decision rules, since the price Π_j occurs only as an irrelevant constant of proportionality. Of course, if the voters were endowed with diverse patterns of votes on the bills then the conclusion in (1) would not hold, and the formulation in terms of decision rules would not be an accurate model of an exchange economy for vote trading, although (2) would remain valid.

The exact meaning of the *decision-rule* formulation depends, therefore, mainly on whether the citizens do or can infer information about the disposition of a bill or about other citizens' preferences from knowledge of the equilibrium prices in the *exchange-economy* game. The solution to this question which is adopted implicitly in the present formulation is to suppose that the citizens in using the market to acquire votes take the prices as fixed data without determining the effects that their actions or other citizens' actions may have on the relative magnitudes of the equilibrium prices. That is, the citizens are modeled as price takers as is customarily done in the economic theory of general equilibrium.[2] Usually this approach must presume a large number of citizens. In Section IV a special case will be analyzed to determine exactly how the equilibrium prices depend upon the citizens' preferences.

II. *Conditions on the Social Choice Process*

Following Arrow's classic treatise [1], we shall develop our general model of an exchange economy for vote trading axiomatically in terms of a set of conditions imposed on the citizens' preferences, the social choice, and (unlike Arrow) the citizens' decision rules. These conditions are intended to confine the model to satisfac-

[1] I am indebted to Robert Aumann and Lloyd Shapley for this observation.

[2] Alternatively, in terms of a Walrasian *tatonnement*, we are supposing a "recontracting" process in which each citizen offers a demand schedule for votes on the various bills as a function of the prices, and by analyzing these schedules a central authority determines a set of equilibrium prices equating supplies and demands. The decision-rule formulation then considers only the demand schedules associated with equilibrium sets of prices.

tory social choices by criteria accepted on prima facie grounds. In part they delimit the scope of reasonable or rational behavior by the citizens, and in part they constrain the social choice produced from the citizens' participation in the collective decision process. In each case they parallel Arrow's development as closely as possible, in the narrower context of vote trading in majority-rule referenda.

1. *The Citizens' Preferences*

In Arrow's general formulation, the set of social states from which the social choice is selected is a primitive datum; here, it is compounded from the bills in the referendum. A social state is a set of bills passed and a complementary set failed, or just the former is a sufficient description. That is, the set S of social states is the set of subsets of the set J of bills proposed, $S=\{x|\phi\subseteq x\subseteq J\}$, where the null set ϕ is the status quo in which all of the bills are failed. Each citizen i is presumed to have a preference ordering R_i among the social states which is complete and transitive. The notation xR_iy means for $x\epsilon S$ and $y\epsilon S$ that the i-th citizen either strictly prefers x to y or is indifferent. The relation of strict preference is P_i, and on $S\times S$ we define the set $R^i=\{(x, y)|xR_iy\}$ for each citizen to characterize his preferences. Further, as was indicated initially, we are interested in the special case in which the citizens see the bills as independent: this is conveyed by the third part of the following condition imposed on the citizens' preferences.

Condition 1: All possible preference orderings of the social states by the citizens are admissible for which for every citizen i and all social states x, y, and z,

(a) either xR_iy or yR_ix or both (Completeness);
(b) xR_iy and yR_iz only if xR_iz (Transitivity);
(c) xR_iy if and only if $(x+z)R_i(y+z)$, where z is disjoint from both x and y (Independence).

Except for (c) this is Arrow's Condition 1′ [1, p. 96]. Part (c) requires that a citizen's preference for the passage of any set of bills does not depend upon the disposition of any other bills.[3] For example, if a voter prefers bill 1 to pass, given that bill 2 will pass (or fail), then he will prefer 1 to pass also given that 2 fails (or passes). This property defines independence of the bills in the present context.

The scope of the admissible set of citizens' preferences precludes dependence of the social choice process upon a more restricted or homogenous set, and forces it to cope with the possibility of diversity, the central problem of social choice.

2. *The Social Choice*

The restrictions on the social choice include, first, the requirement in the present model that it is to be selected by majority-rule after vote trading in an exchange economy.

Condition 2: The social choice is the social state $x^0\epsilon S$ for which $x^0=\{j|\ \sum_{i\epsilon I}d_j^i>0\}$, where d^i is the i-th citizen's decision rule.

The inequality is the condition for passage that the affirmative votes exceed the negative votes, here scaled by the equilibrium price. Second, it is required that the social choice be Pareto optimal, essentially a weak form of Arrow's condition of the positive association of social and individual values.

[3] Condition 1 does not exclude the Condorcet paradox, the case of cyclical majorities among social states which is fundamental to Arrow's general construction, in contrast to the several proposals that have been advanced by Duncan Black [3], Kenneth Arrow [1], Amartya Sen [14], P. K. Pattanaik [11] and others to replace Arrow's Condition 1 by the Condition of Single-Peakedness (Arrow's Condition 1″) or one of its generalizations.

Condition 3: If xR_iy for every citizen i, and $xP_{i'}y$ for some citizen i', then the social choice x^0 is not y.

Arrow's Conditions 4 (Citizens' Sovereignty) and 5 (Nondictatorship) are essentially assured by our Condition 2 (Majority Rule). His Conditions 2 (Positive Association of Social and Individual Values), 3 (Independence of Irrelevant Alternatives), and 4 were used to derive his Consequence 3 or Condition P (Pareto Optimality) in a slightly weaker form than our Condition 3 although our conditions do not imply his (especially Condition 3, [1, p. 97]). The desirability of a social choice process achieving Pareto optimality is universally appreciated. Nevertheless, one should not underestimate the force of requiring it in every instance. In game-theoretic terms, it enforces a cooperative solution and, therefore, excludes an enormous range of decision rules that in practice a citizen would find to be of strategic value. For example, it excludes virtually all rules derived from probability judgments about others' actions, such as hypothesized by Coleman [5].

3. *The Citizens' Decision Rules*

Among the three conditions so far imposed, Conditions 1 and 3 are nearly direct analogues of Arrow's Conditions 1′ and P. Our Condition 2 (majority rule with vote trading), however, is essentially ambiguous in relation to Arrow's construction. Although it imposes a special form on the social choice process, it leaves the citizens' decision rules unspecified to a large degree; indeed, I would guess that a great variety of decision rules are compatible with Pareto optimality as required in Condition 3. At this point in our analysis, an extension of Arrow's approach would impose further normative conditions on the social choice, corresponding to Arrow's Conditions 2–5. An essentially different approach will be employed here, however. Several conditions will be imposed to circumscribe the set of admissible decision rules, in part to define rational behavior by the citizens, and in part to exclude any dependence of the social choice process upon idiosyncratic behavior by one or several of the citizens. In this way, we confine the axiomatic development to essentially descriptive statements about the citizens and the social choice process except for the normative content isolated in Condition 3 (Pareto Optimality). However, as a referee of this paper has pointed out, to whatever extent the conditions on the citizens' decision rules limit behavior they are potentially normative, in contrast to Arrow's approach in which normative or rationality conditions are imposed only upon the social choice.

The conditions to be imposed on the admissible decision rules are of two kinds: the first three ensure a sufficient diversity in the admissible set, and the fourth restricts the admissible set.[4] In the following, a citizen's decision rule is regarded as a function of all the citizens' preference orderings, say $d^i(R^1, \cdots, R^m)$.

Definition: The set of admissible decision rules is *symmetric* if for any two citizens i and i', $d^i(\cdots R^i \cdots R^{i'} \cdots)$ is an admissible decision rule for the i-th citizen only if $d^{i'}(\cdots R^{i'} \cdots R^i \cdots) = d^i(\cdots R^i \cdots R^{i'} \cdots)$ is admissible for the i'-th citizen.

That is, a symmetric set is one in which two citizens might act the same in the same situation: a symmetric set merely allows that what is rational for one citizen may be rational for another.[5] Symmetry

[4] Note that the first two hypothesize variations in the citizens' preferences, as in Arrow's Conditions 2–5, although we need to consider only permutations of entire preference orderings, whereas the fourth hypothesizes a variation in the set of social states, as in Arrow's Condition 1.

[5] Another definition of symmetry requires that $d^i(R^{\alpha(1)}, \ldots, R^{\alpha(m)}) = d^{\alpha(i)}(R^1, \ldots, R^m)$ is admissible, where α is a permutation of the citizens' indices.

of the admissible set is a very weak condition restricting idiosyncratic behavior by the citizens.[6]

> *Definition:* The set of admissible decision rules is ***independent*** of the labeling of the bills if for every citizen i, $d^i(R^1, \cdots, R^m)$ is admissible only if for all $j \epsilon J$ $d^i_{\alpha(j)}(Q^1, \cdots, Q^m)=d^i_j(R^1, \cdots, R^m)$ is admissible, where α is a permutation of the indices of the bills which takes R^i into Q^i for all $i \epsilon I$.

To require that the admissible set be independent is merely to allow that a citizen need not be deterred in his strategy by a change in the order of presentation of the bills. Independence is another very weak condition restricting idiosyncratic behavior by the citizens.

> *Definition:* The set of admissible decision rules is ***neutral*** if for every bill j, $d^i_j=0$ is admissible whenever citizen i is indifferent between social states $\{j\}$ and ϕ.

Neutrality allows a voter to be neutral on bills about which he is indifferent, rather than forcing him to commit some of his voting power one way or the other. The fourth property of the admissible set that we shall require is admittedly stronger in its behavioral content, since it presumes an element of rationality often missing in practice. Above, we described the status quo as the failure of all of the bills. However, alternatives may be constructed as follows: let a subset y of the proposed bills be considered passed, and in their place introduce new bills designed to defeat or recall the bills in y. In this case, the status quo ϕ is transformed into y, and any set of bills $x \epsilon S$ is transformed into $x-x\cap y+(J-x)\cap y$; in particular, y is transformed into the new status quo labeled ϕ. We shall require that such variations in the status quo be irrelevant to the social choice by restricting the admissible decision rules appropriately.

> *Definition:* The set of admissible decision rules is *stable* if for any transformation $T(y)$, $y \epsilon S$, of the set of social states which takes each $x \epsilon S$ into $(x-x\cap y+(J-x)\cap y) \epsilon S$, and which transforms the citizens' preferences correspondingly, each citizen i's decision rule $d^i=(d^i_j)$ is transformed into $\delta^i=(\delta^i_j)$, where $\delta^i_j=d^i_j$ for $j \not\epsilon y$ and $\delta^i_j=-d^i_j$ for $j \epsilon y$.

Stability requires that a change in the status quo merely reverses the citizens' affirmative and negative declarations for the bills in y considered passed a priori. Like Arrow's Condition 1, stability hypothesizes a variation in the set of social states: I think any deep analysis of social choice must do so at some point.

To summarize, we gather together the conditions imposed on the citizens' decision rules.

> *Condition* 4: The admissible set of citizens' decision rules is symmetric, independent of the labeling of the bills, neutral, and stable.

It should be noted that none of the four conditions imposed on the admissible decision rules hypothesizes citizens' preferences violating Condition 1, in particular (c) of Condition 1.

III. *A Characterization of Individualistic Decision Rules*

Conditions 1–4 provide a general model of social choice by majority rule among independent bills when there is an exchange economy for vote trading. In this model, the social choice is constrained by the method of majority rule, the normative criterion of Pareto optimality, and

[6] In the case of a representative assembly one might test for symmetry by examining whether, if two legislators interchanged constituencies, they could feasibly adopt each other's voting strategies; e.g., symmetry might be violated if seniority or party affiliations were major factors.

the admissibility conditions imposed on the citizens' decision rules. Unfortunately, the decision rules consistent with these conditions have not been determined in general. The purpose of our subsequent analysis is to characterize completely the citizens' decision rules in an important special case.

The special case we shall consider is the instance of a referendum as one would normally occur in practice; namely, in which there are many citizens (actually it will be sufficient that the number of citizens exceed the number of bills, $m \geq n$) acting independently without opportunities to know or take advantage of others' preferences or voting strategies. Note that this situation is consistent with our earlier presumption of a price-mediated market for vote trading accomplished anonymously. The main result will show that the conditions imposed require that a citizen's voting strategy be *sincere*, in the sense that his allocation of his voting power among the bills must constitute a cardinal measure of his preferences for the bills.

Definition: An additive utility representation of the i-th citizen's preference ordering R_i of the social states is a set of real numbers $\{u_{ij} | j=1, \cdots, n\}$, one for each bill, such that xR_iy if and only if $\sum_{j\epsilon x} u_{ij} \geq \sum_{j\epsilon y} u_{ij}$ for all $x \epsilon S$, $y \epsilon S$.

We assert that each citizen's decision rule must be such an additive utility representation of his preferences, unless possibly he is indifferent among some social states.

The special case is defined by the following additional condition, requiring that each citizen's decision rule does not depend upon others' preferences.[7]

Condition 5: For each citizen i, if $Q^i = R^i$ then $d^i(Q^1 \cdots, Q^m) = d^i(R^1, \cdots, R^m)$ for all admissible preference orderings of the citizens.

The result is the following.

Theorem: If $m \geq n$ and any citizen i is not indifferent between any two social states, then Conditions 1-5 imply that his decision rule is an additive utility representation of his preferences.

The proof must show that xP_iy if, and only if, $\sum_{j\in x} d_j > \sum_{j \in y} d_j^i$. The *if* is trivial if the converse has been demonstrated. Also, Condition 1(c) shows that it suffices to suppose that x and y are disjoint. Indeed, we can assume $y = \phi$ since otherwise we apply the transformation $T(y)$ defined in the definition of stability; viz., xP_iy is transformed to $(x+y)P_i\phi$ and, therefore, $\sum_{j\in x} d_j^i + \sum_{j \in y}(-d_j^i) > 0$ as required. Hence, suppose (for convenience) $i=1$, and that $xP_1\phi$ but $\sum_{j \in x} d_j^1 \leq 0$. Let k be the number of bills in x and note that $k \leq n \leq m$ by hypothesis. Let $\alpha_1, \cdots, \alpha_k$ be the k cyclic permutations of the indices of the bills in x (the others remaining unchanged), where α_1 is the identity. Condition 1 allows us to consider the case in which α_i maps R^1 into R^i for $i \leq k$; and in which for $i > k$, $\{j\}$ and ϕ are indifferent for each $j \in x$. The neutrality condition allows us to suppose that $d_j^i = 0$ for $i > k$, $j \in x$. Further, for $i \leq k$ we can suppose that $d_j^1(R^1) = d^1_{\alpha_i(j)}(R^i) = d^i_{\alpha_i(j)}(R^i)$ for all $j \in x$, where the first equality is allowed by the condition of independence and the second

[7] In view of the remarks made at the end of Section I, this condition is even more restrictive than it may at first appear. For, if in an *exchange-economy* game a citizen's choice of the numbers of votes he acquires on the various bills may depend only on his own preferences and in an arbitrary way upon the equilibrium prices, Condition 5 would be violated because the equilibrium prices will generally depend upon all the citizens' preferences. In an *exchange-economy* game, therefore, Condition 5 asserts that the valuation (at the equilibrium prices) of the votes a citizen acquires on any bill does not depend upon the prices. Given the social choice process assumed in Condition 2 this is entirely reasonable, but it is certainly not necessary. In Section IV we shall see that the equilibrium prices do not convey any information about the distribution of pro and con sentiments about a bill among the citizens, which is at least consistent with Condition 5.

by symmetry. Consequently,

$$\sum_{i\in I} d_j^i = \sum_{j\in x} d_j^1 \leq 0$$

for all $j\in x$ and, therefore, by Condition 2 all the bills in x will be defeated. But this violates Pareto optimality; since, for all $i\leq k$ the i-th citizen strictly prefers (using Condition 1(c)) $x+x^0$ to x^0 for any social choice x^0 selected from the other bills not in x, and the other citizens are indifferent. Thus, the original supposition that $\sum_{j\in x} d_j^1 \leq 0$ must have been false, as was to be shown, and the proof is complete. This proof is roughly a variation on the theme of cyclical majorities employed by Arrow.

The theorem is rather remarkable in one technical aspect; specifically, it asserts that the wide range of admissible citizens' preference orderings required in Condition 1 is not compatible with the conditions 2–5. Some time ago, Bruno de Finetti [7] and L. J. Savage [12] raised the question of whether Condition 1 implies the *existence* of an additive utility representation, but the issue was resolved negatively by Charles Kraft, John Pratt, and A. Seidenberg [8] in 1959 with a counter-example for $n=5$ with strict preferences. (A simpler proof is by Dana Scott [13].) They proved that the following is the necessary and sufficient condition for the existence of an additive utility representation of the i-th citizen's preferences.

Condition 6: If $\{(x^k, y^k)\,|\,k=1, \cdots, K\}$ is a finite sequence of elements from R^i, and if for each $j\in J$ the numbers of x's and y's containing j are equal, then $(y^k, x^k)\in R^i$ for all $k=1, \cdots, K$.

This condition implies Conditions 1(b) and 1(c) in the presence of 1(a), as can be seen by considering the sequence $\{(x, y), (y, z), (z, x)\}$ for $K=3$ to establish Condition 1(b), and the sequence $\{(x, y), (y+z, x+z)\}$ for $K=2$ to establish Condition 1(c). The Kraft, Pratt, and Seidenberg [8] result yields the following additional result in the present context.

Corollary: If $m\geq n$ and any citizen i is not indifferent between any two social states, then Conditions 1–5 imply that the i-th citizen's preferences satisfy Condition 6.

That is, Condition 6 is a necessary condition for the existence of the additive utility representation asserted in the theorem. Note that since the citizen's preferences are strict, the hypothesis and conclusion of Condition 6 can both be true only if $x^k=y^k$ for all $k=1, \cdots, K$. The significance of this Corollary is that it demonstrates that the Conditions 2–5 imposed on the social choice and the citizens' decision rules constitute a restriction on the admissible citizens' preferences.

IV. *Analysis and Discussion*

In terms of our original description of vote trading, the theorem above asserts that if $m\geq n$, all preferences are strict, and Conditions 1–5 are satisfied, then there must exist an additive utility representation $\{u_{ij}\}$ of each citizen i's preferences such that for each bill j,

$$u_{ij} = \mu_i\sigma_{ij}\,\Pi_j\,V_{ij} \tag{1}$$

(where $\sigma_{ij}u_{ij} = |u_{ij}|$)

for some constant $\mu_i>0$. Multiplying by σ_{ij} and summing this relation among the bills, and assuming a budget equality $\sum_{j\in J}\Pi_j V_{ij}=1$, yields $\mu_i=\sum_{j\in J}|u_{ij}|$. That is, μ_i is a scale parameter measuring the utility difference or *gain* between the worst and best social states for the i-th citizen. Of course, we can scale this gain to unity for each citizen, since utility

scales are arbitrary. Similarly, summing this time among all citizens and using the condition $\sum_{i\in I} V_{ij}=m$ yields $\Pi_j=(1/m)\sum_{i\in I}|u_{ij}|$; that is, the price of a vote on the j-th bill is proportional to the sum of the *stakes* in the bill.[8] Finally, the j-th bill passes if and only if

$$\sum_{i\in I} \sigma_{ij} V_{ij} > 0,$$

or equivalently $\sum_{i\in I} u_{ij}>0$.

The way in which this model violates the Independence of Irrelevant Alternatives is now apparent. Suppose that in a particular instance the n-th bill is defeated. The Independence of Irrelevant Alternatives requires that if the n-th bill had never been proposed for the referendum then the same social state would have resulted. However, if among the citizens there were diverse stakes in the n-th bill then this is not assured at all; for, after eliminating the n-th bill, rescaling of the citizens' gains to unity will produce diverse effects on their stakes in any remaining bill, possibly to the extent of changing its disposition. Thus, the mere proposal of a bill, even if subsequently defeated, may affect the passage or failure of other bills since it will require allocations of the citizens' voting power in varying amounts depending upon their respective stakes in it.

Alternatively, one can construe the imputed equality among the citizens' gains as an interpersonal comparison of utilities, which is prohibited by the Independence of Irrelevant Alternatives.

As an example, suppose there are three bills in a referendum presented to three citizens who rank the social states as follows (Table 1):

[8] Note that the price of a vote on a bill is independent of the distribution of pro and con sentiments among the citizens.

TABLE 1

SOCIAL STATE (BILLS PASSED)	RANK ORDERINGS Citizens (i): 1	2	3
ϕ	3	2	8
1	4	4	4
2	1	6	7
3	7	1	6
1,2	2	8	3
1,3	8	3	2
2,3	5	5	5
1,2,3	6	7	1

In this case the following decision rules are admissible (Table 2):

TABLE 2—DECISION RULES (d_j^i)

CITIZEN (i)/Bill (j):	1	2	3
1	−1/12	3/12	−8/12
2	−3/12	−8/12	1/12
3	8/12	1/12	3/12

And the votes as cast will be as follows, with the indicated dispositions of the bills and the prices of votes (Table 3):

TABLE 3—VOTES

Bill (j)	1	2	3
For	2	1	1
Against	1	2	2
Price	1/3	1/3	1/3
Disposition	Pass	Fail	Fail

In this example, bill 1 passes. However, if bills 2 and 3 are eliminated from consideration then bill 1 must fail since a majority of citizens ($i=1$ and 2) opposes it and no vote trading is possible.

It is worth noting also that in the above example direct balloting among alternative social states, as in Arrow's characterization of logrolling, would yield an intransitivity among the triplet of social states (ϕ, 1, 2), among several others.

V. *Generalization*

The special case of individualistic decision rules, considered above in Sections III and IV, is obviously too restrictive to be of great value in the analysis of realistic voting situations. It is worthwhile, therefore, to indicate briefly the directions in which the present results may be extended. Elsewhere, using the theory of cooperative games without side payments, I have begun the analysis of the general case posed by Conditions 1–4 in Section II.[9]

The most obvious fault with Condition 5 is the unrealistic exclusion of any dependence of a citizen's decision rule upon the set of equilibrium prices, say of the form $d^i(R^i; \Pi)$ where $\Pi = (\Pi_j)$. Indeed, it is evident that regardless of the practical limitations on a citizen's ability to know others' preferences and strategies, he must know the prices if the model is to conform to an exchange economy: certainly a recontracting process (cf. fn. 3) would presume this information. Also, the situation constructed to obtain the contradiction in the proof of the theorem is consistent only with the instance of equal prices for votes on all bills, as in the example in Section IV.

An appealing way of analyzing price dependencies is to enforce the Weak (or Strong) Axiom of Revealed Preference, but I have not obtained any sharp characterizations from this approach. A less appealing but more successful approach has been the following. While weakening Condition 5 to allow price dependencies, strengthen Condition 4 by appending the following condition.

Condition 7: For each subset of bills $H \subset J$, $\phi \neq H \neq J$, and each scalar λ_i such that $0 < \lambda_i \sum_{j\in H} |d_j^i(R^i; \Pi)| \leq 1$, there exists an admissible preference ordering Q^i for the i-th citizen such that R^i and Q^i agree on H and $d_j^i(Q^i; \Pi) = \lambda_i d_j^i(R^i; \Pi)$ for all $j \in H$.

That is, by varying a citizen's preferences for bills not in H, one can vary arbitrarily the proportion of his budget allocated to the bills in H without altering the distribution among the bills in H. Of course, the reasonableness of this condition depends upon Condition 1(c), and in any case it is a very strong condition. Together with Conditions 1–4, and 5 as modified, Condition 7 implies the following weaker version of the theorem in Section III. Suppose $n \leq m$ and that for some $x \in S$, $x \neq J$, $xP_1\phi$. Let $\beta = \{\beta_i | i \in I\}$ be any set of permutations on the indices of the bills in x for which β_1 is the identity. Then, for each such β there exists a set of scalars $\xi = \{\xi_j | j \in x, \xi_j \geq 0\}$ such that

$$\sum_{j\in x} \xi_{\beta_i(j)} d_j^1(R^1; \Pi^i) \geq 0 \qquad (> \text{if } i = 1)$$

for all $i \in I$, where β_i maps Π^i into Π. The proof, which is omitted here, employs the duality theory of linear inequalities.[10] A little consideration will show that the effect of this result is to restrict the variations in the decision rule arising from price effects to a fairly narrow range around the decision rule predicted in the case of no price dependencies: the more numerous the bills, the tighter the restriction. If there are no price dependencies, then $\xi_j = 1$ for all $j \in x$ suffices and we obtain our earlier results once again. This appears to be nearly the strongest result that can be obtained by considering only permutations; hence, another approach may be more effective in further analyses.

[9] Robert Wilson, [15] and [16]; the former was presented at the Social Choice Conference, University of Pittsburgh, September 9–12, 1968, and will be included in the forthcoming volume of its Proceedings edited by Bernhardt Lieberman and R. Duncan Luce.

[10] The proof follows closely the lines of the proof of the earlier theorem.

REFERENCES

1. K. J. ARROW, *Social Choice and Individual Values*, 2nd ed. New York 1963.
2. A. BENTLEY, *The Process of Government*. Evanston 1908, 1935, 1949, pp. 370–371.
3. D. BLACK, *The Theory of Committees and Elections*. Cambridge 1958.
4. J. BUCHANAN AND G. TULLOCK, *The Calculus of Consent*. Ann Arbor 1962.
5. J. COLEMAN, "The Possibility of a Social Welfare Function," *Am. Econ. Rev.*, Dec. 1966, *56*, 1105–22.
6. ———, "Reply," *Am. Econ. Rev.*, Dec. 1967, *57*, 1311–17.
7. B. DE FINETTI, "La 'Logica del plausible' secondo la concezione de Polya," *Atti della XLII Riuione della Societa Italiana per il Progresso delle Scienze*, 1949, 1–10.
8. C. KRAFT, J. PRATT, AND A. SEIDENBERG, "Intuitive Probability on Finite Sets," *Annals Math. Stat.*, 1959, *30*, 408–419.
9. D. C. MUELLER, "Comment" on [5], *Am. Econ. Rev.*, Dec. 1967, *57*, 1304–11.
10. R. E. PARK, "Comment" on [5], *Am. Econ. Rev.*, Dec. 1967, *57*, 1300–1304.
11. P. K. PATTANAIK, "A Note on Democratic Decision and the Existence of Choice Sets," *Rev. Econ. Stud.*, Jan. 1968, *35*(*1*), 1–9.
12. L. SAVAGE, *The Foundations of Statistics*. New York 1954.
13. D. SCOTT, "Measurement Structures and Linear Inequalities," *J. Math. Psych.*, 1964, *1*, 233–217.
14. A. K. SEN, "A Possibility Theorem on Majority Decisions," *Econometrica*, Apr. 1966, *34*, 491–499.
15. R. B. WILSON, "A Game-Theoretic Analysis of Social Choice," Paper 145, Sept. 1968, Stanford Business School.
16. ———, "A Class of Solutions to Voting Games," Paper 156, Oct. 1968, Stanford Business School.

Price Expectations and the Phillips Curve

By ROBERT E. LUCAS, JR. AND LEONARD A. RAPPING*

Until recently, there have been two main approaches to the rationalization of the observed negative correlation between the unemployment rate and the rate of inflation (the Phillips curve)[1]. Phillips [19] and Lipsey [8] postulated a competitive adjustment mechanism, where the rate of change in money wages is negatively related to excess supply in the labor market, with the latter quantity measured by the unemployment rate. A second view, expressed by Eckstein and Wilson [3] and Perry [17] emphasizes a collective bargaining mechanism with unemployment rates and other variables (like profit rates) measuring union bargaining power over (or employer resistance to) increases in money wages.[2] Both approaches, as implemented empirically by these and other authors, imply a permanent and rather discouraging tradeoff between inflation and unemployment. The results have led to two important policy conclusions: first, that sustained inflation is both necessary and sufficient to the sustained maintenance of low unemployment rates; and, second, that policies (like wage-price guideposts) which may shift the Phillips curve, are an important ingredient of national economic policy.

Recently, several authors have sought to elaborate different theories underlying the Phillips curve, emphasizing the role of expectations in labor markets either in addition to, or in place of, the bargaining mechanisms referred to above.[3] These approaches, while differing considerably in detail, all suggest that the inflation-unemployment trade-off is a short run phenomenon, or that sustained inflation will make *no* contribution to the permanent lowering of unemployment rates. In view of the policy implications drawn from the original approaches, as sketched above, it is clearly crucial to attempt to discriminate empirically between what may be called the "bargaining mechanism approach" and the "expectations approach" to a theory of the Phillips curve.

The objective of this paper is to review a particular version of the "expectations approach" (our own), to test this version on

* The authors are associate professors at Carnegie-Mellon University. They wish to thank T. W. McGuire and E. Phelps who commented on an earlier draft of this paper.

[1] Throughout this paper the term Phillips curve is used interchangeably to refer either to a price change-unemployment relationship or to a money wage change-unemployment relationship.

[2] These remarks are not intended to suggest that the references cited contain a well articulated theory of wage-employment determination in an economy dominated by collective bargaining. We have been unable to find such a theory anywhere in the literature. We have, however, attempted to construct a rationalization of the Phillips curve along these lines ourselves—an attempt which ran into two difficulties we found insurmountable. First, why should "money illusion" characterize the outcome of collective bargaining if the individuals and employers involved are free of it? Second, even if one had an adequate theory of wage determination in a single, unionized sector, what implications will this theory have for aggregate wage and employment determination? (This later point is developed more fully in [9].) In part, the union impact on aggregate wages depends on the wage interconnections between the unionized sector and the rest of the economy. In several recent studies, McGuire and Rapping, [11], [12], [13], [14] argue that most of the evidence on "spillovers" from unionized to nonunionized sectors is either ambiguous or irrelevant to *absolute* wage determination. For an application of a bargaining based Phillips curve to the entire economy (as opposed to the manufacturing sector studied in [3], and [18]) see Tella and Tinsley [21].

[3] We have in mind our own previous study, [9], as well as Phelps [18], Friedman [4], Meltzer [10], and Mortenson [15].

United States time series for 1904–65, and to use this theory as a framework for isolating short and long run unemployment-inflation trade-offs. Theory, tests, and conclusions are given in sections I, II, and III respectively.

I. *An Expectations Theory of the Phillips Curve*

In an earlier study[4] we postulated an aggregate labor supply function of the form:

$$\begin{aligned} & ln\ (N_t/M_t) \\ (1)\qquad & = \beta_0 + \beta_1\, ln\ (w_t/w_t^*) + \beta_2\, ln\ (w_t^*) \\ & \quad - \beta_3\, ln\ (P_t^*/P_t), \end{aligned}$$

where N_t is total man-hours supplied, M_t is population, w_t is the current real wage (W_t/P_t), and w_t^* and P_t^* are "permanent" or "normal" real wages and prices respectively. In [9], eq. (1) is generated by a two-period Fisherian analysis of the household's labor-leisure choice problem, which implies that β_1 and β_3 are positive, while β_2 may have either sign. Estimates reported in [9] indicate that β_2 is approximately zero, or that the long run labor supply schedule is approximately vertical. On the other hand, it was found that labor supply is highly responsive to changes in w_t/w_t^*, the ratio of current to normal wages, and to the anticipated inflation rate, $ln(P_t^*/P_t)$. Thus the short and long run behavior of labor supply, as predicted by our theory and confirmed in our tests, is as indicated in Figure 1.[5]

Combining this supply model with a labor demand function, *and* assuming that labor markets are cleared each period, one has a complete aggregate labor market model (once the formation of w_t^* and P_t^* is described). Completing the model in this way,

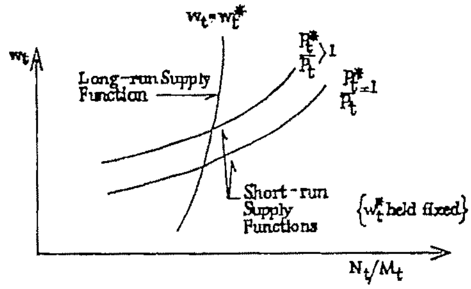


FIGURE 1

the interpretation of all or part of measured unemployment as an excess supply in the usual sense is precluded. We sought, then, an alternative hypothesis as to what people mean when they answer "yes" to the Census Bureau question: "Are you actively seeking work?" We hypothesized that respondents to this question take it to mean: "Are you seeking work at your *normal* wage rate?" Thus we assumed that the labor force as measured by the employment survey consists of employed persons plus those who are searching for work at what they regard as their normal wage rate (or, roughly equivalently, in their normal occupation).

To define unemployment (in our sense) more precisely, we define the *normal labor supply*, N_t^*, to be the quantity which would be supplied when the actual wage rate, w_t, equals the anticipated normal wage, w_{t-1}^*, and when actual prices, P_t, equal normal prices, P_{t-1}^*. Under these conditions, (1) becomes:

$$\begin{aligned} (2)\qquad ln\ (N_t^*/M_t) &= \beta_0 + \beta_1\, ln\ w_{t-1}^* \\ &\quad + (\beta_2 - \beta_1)\, ln\ w_t^* - \beta_3\, ln\ (P_t^*/P_{t-1}) \end{aligned}$$

Subtracting (1) from (2) gives:

$$\begin{aligned} (3)\qquad ln\ (N_t^*/N_t) &= \beta_1\, ln\ (w_{t-1}^*/w_t) \\ &\quad + \beta_3\, ln\ (P_{t-1}^*/P_t). \end{aligned}$$

Because $ln\ (N_t^*/N_t) \sim (N_t^* - N_t)/N_t^*$, the left side of (3) has the dimensions of an unemployment rate, but it will differ from the measured unemployment rate, U_t, for two reasons. First, teenagers and women may

[4] Since considerable space is devoted to explaining equation (1) in our earlier paper [9], treatment here will be sketchy.

[5] An anticipated rise in the price of future goods induces, among other effects, a substitution of current leisure for future goods consumption—hence a decrease in current labor supply (other prices, including the nominal interest rate, remaining fixed).

fail to report themselves as unemployed and, secondly, N_t^* is defined in terms of a representative household and so does not allow for the frictional component of measured unemployment. Since there is reason to think that the frictional and nonfrictional components of total unemployment vary together, we assume that U_t and $ln\,(N_t^*/N_t)$ are linearly related:

$$(4)\quad U_t = g_0 + g_1\, ln\,(N_t^*/N_t), \quad g_0, \quad g_1 > 0.$$

Combining (3) and (4):

$$(5)\quad \begin{aligned} U_t = g_0 &+ g_1\beta_1\, ln\,(w_{t-1}^*/w_t) \\ &+ g_1\beta_3\, ln\,(P_{t-1}^*/P_t). \end{aligned}$$

It remains only to link the unobserved, normal wages and prices to observable, actual values. In our original model we assumed that both normal variables are adaptively calculated by households from actual values, with the same reaction parameter λ, $0<\lambda<1$, or that:

$$(6)\quad ln\,(w_t^*) = \lambda\, ln\,(w_t) + (1-\lambda)\, ln\,(w_{t-1}^*),$$

and:

$$(7)\quad ln\,(P_t^*) = \lambda\, ln\,(P_t) + (1-\lambda)\, ln\,(P_{t-1}^*).$$

Using standard methods to reduce the system of difference equations (5)–(7) to a single difference equation in U_t, we obtain:[6]

$$(8)\quad \begin{aligned} U_t = \lambda g_0 &- g_1\beta_1\, ln\,(w_t/w_{t-1}) \\ &- g_1\beta_3\, ln\,(P_t/P_{t-1}) \\ &+ (1-\lambda)U_{t-1}. \end{aligned}$$

The reader should note that when going from (5) to (8) the coefficients of the wage and price terms change sign because in (6) and (7) we assume an "elasticity of expectations" less than unity (when current prices or real wages rise, suppliers anticipate an eventual return to normalcy). Since (8) relates unemployment to the inflation rate (*celeris paribus*) it is a Phillips curve in the broad sense.[7] In the next section, estimates of (8) for 1904–65, and various subperiods, are reported.

As with the other empirical Phillips curves, (8) implies a long run as well as a short run inflation-unemployment trade-off, with the consequent promise of a *permanent* decrease in unemployment if the economy is willing to tolerate a sustained inflation. This long run trade-off is, however, built into (8) in a transparent way by the expectations hypotheses (6) and (7). Thus the reason (8) offers a long run trade-off lies in the assumption of an unreasonable stubborness on the part of households: if a sustained inflation policy is pursued by the government, households following (7) will continue *forever* to underpredict future prices.

There is no entirely satisfactory way to remedy this deficiency in the theory within the framework of adaptive expectations. Any forecaster predicting future prices as a fixed function, however complicated, of past prices can be systematically fooled by a clever opponent manipulating the actual series at will. But since there is little reason to believe that the government *systematically* manipulates prices, it may be worthwhile to examine adaptive schemes which, unlike (6) and (7), permit a short run Phillips curve without deciding the question of the existence of a long run curve a priori.

Focusing on the anticipation of prices rather than real wages, a natural generalization of (7) is the hypothesis that $ln(P_t^*)$ is a "rational distributed lag function" of past actual values,[8] or that:

[6] See note 10.

[7] Since knowledge of the rates of change of any two of the three variables w_t, P_t and w_tP_t (money wages) implies knowledge of the rate of change of the third, (8) can be rewritten in two other ways. In particular, one may relate unemployment to money wage changes, as is more conventionally done.

[8] The term "rational distributed lag function" is from Jorgenson [6], where the word "rational" applies to the form of the generating function of the distributed lag. Jorgenson shows that a distributed lag has a rational generating function if and only if it can be ex-

$$
\begin{aligned}
ln\,(P_t^*) &= b_0\, ln\,(P_t) + b_1\, ln\,(P_{t-1}) \\
&\quad + \cdots + b_r\, ln\,(P_{t-r}) \\
&\quad + a_1\, ln\,(P_{t-1}^*) + \cdots \\
&\quad + a_s\, ln\,(P_{t-s}^*).
\end{aligned} \tag{9}
$$

As in (7), we impose the condition:

$$b_0 + \cdots + b_r + a_1 + \cdots + a_s = 1, \tag{10}$$

so that a proportional change in *all* past prices would imply a change in P_t^* of the same proportion.[9] We assume as well that (9) is stable.

Equations (5), (6) and (9) form a system of three linear difference equations in U_t, $ln(P_t^*)$ and $ln(w_t^*)$ with forcing variables $ln(P_t)$ and $ln(w_t)$, which may, using the restriction (10), be reduced to a difference equation in U_t of the form:[10]

$$
\begin{aligned}
U_t &= a + c_0\Delta\, ln\,(P_t) + \cdots \\
&\quad + c_k\Delta\, ln\,(P_{t-k}) + d_0\Delta\, ln\,(w_t) \\
&\quad + \cdots + d_n\Delta\, ln\,(w_{t-n}) \\
&\quad + e_1U_{t-1} + \cdots + e_mU_{t-m},
\end{aligned} \tag{11}
$$

where $\Delta x_t = x_t - x_{t-1}$. The coefficient c_0 of the current inflation rate in (11) is equal to $g_1\beta_1(b_0-1)$ so that the existence of a downward sloping short run Phillips curve is equivalent to the condition $b_0<1$ (an "elasticity of expectations" less than unity). The slope of the long run Phillips curve is obtained by summing the coefficients $c_0 + \cdots + c_k$ and dividing by one minus the sum of the lagged unemployment terms, $e_1 + \cdots + e_m$.

For $k=0$, (11) implies that there will be a long run Phillips curve if and only if there is a short run Phillips curve, both cases occurring whenever $c_0<0$. For $k\geq 1$, however, it is entirely possible that $c_0<0$ and simultaneously $c_0 + \cdots + c_k = 0$, so that the slope of the long run curve is zero.

In the next section, we report estimates of the coefficients of (11) for the case $k=2$, $n=1$, $m=2$—the values which, after some experimenting, appeared to yield the "best" results.[11] Using this case as a maintained hypothesis, we also test the hypotheses $c_0+c_1+c_2=0$ and $d_0+d_1=0$. In the testing and estimation, we assume that (11) is disturbed by an error term ϵ_t, where $\{\epsilon_t\}$ is a sequence of independent identically distributed random variables with finite variance σ^2 and mean 0, independent of contemporaneous values of w_t and P_t. Under these conditions, and provided (11) is stable,[12] the estimated coefficients will be consistent and asymptotically normal, and certain Chi-square tests (described in detail below) will be approximately valid.

II. *Empirical Results.*

In this section we report tests of (8) and (11), on data series covering the period 1900–65. For the years 1900–60, the unemployment series used is from Lebergott [7, p. 512] and for 1961–65, reported census data are used [1, p. 236]. The price series is the Consumer Price Index taken

pressed in the form (9), which he calls its "final form". This usage should not be confused with "rational expectations" as defined by Muth in [16]; the two concepts are unrelated.

[9] The formulation (9) is sufficiently general to include the best known expectations hypotheses. For example, to obtain (7), let $b_0=\lambda$ and $a_1=1\text{-}\lambda$. For expectations of inflation which are adaptive on the *rates of change* of prices (and hence extrapolative on price levels) set $b_0=1+\lambda$, $b_1=-1$, and $a_1=1\text{-}\lambda$, and other a_i, b_j equal to zero. (This example also shows that one would not wish to impose the restriction $b_i\geq 0$ on (9)). Any convex combination of regressive (as in (7)) and extrapolative expectations can then be formed in an obvious way.

[10] This reduction is most easily carried out by rewriting (5), (6) and (9) in terms of current (time t) values of the variables, using the lag operator E defined by $Ex_t=x_{t-1}$. Then to obtain (11), one multiplies both sides of (5) by $[1\text{-}a_1E\text{-}\ldots\text{-}a_sE^s]$ $[1\text{-}(1\text{-}\lambda)E]$ and substitutes from (6) and (9) to eliminate terms involving w_t^* and P_t^*. Then applying (10) to the resulting expression yields (11). The parameters k, n and m of (11) are found to be $k=max(r,s)-1$, $n=s$, and $m=s+1$. Similarly, the constants c_i, d_j, e_q are determinable functions of $b_0, \ldots, b_r, a_1, \ldots, a_s, \lambda, g_1, \beta_1$ and β_2.

[11] In terms of the parameters of (9), this case corresponds to $r=3$ and $s=1$.

[12] The stability of (11) follows from the assumed stability of (6) and (9).

TABLE 1—ESTIMATES FOR MODEL BASED ON EQUATIONS (5), (6) AND (7).[a]

Variable / Time Period	Constant	$\Delta ln(P_t)$	$\Delta ln(w_t)$	(U_{t-1})	R^2(Adj)	DWS
1904–65	2.32 (.57)***	−25.28 (5.82)***	−21.72 (10.18)*	.822 (.054)***	.832	1.44
1904–29	4.04 (.95)***	−10.84 (6.84)	−29.75 (12.21)**	.322 (.176)	.370	1.67
1930–45	2.50 (1.40)	−70.80 (8.73)***	4.19 (22.45)	.818 (.063)***	.953	2.21
1946–65	4.15 (1.37)***	−13.80 (10.23)	− 4.51 (12.74)	.249 (.208)	.180	1.73
1930–65 from [9]	4.2 (1.0)***	−59 (8)***	−41 (24)	.80 (.05)***	.925	1.50

* } one tail significance at { .025
** } one tail significance at { .01
*** } one tail significance at { .005

[a] The unemployment rate is measured as a percentage while the Δln variables are on the order of magnitude of .01. The effect of a one percentage point increase in the rate of inflation on the percentage unemployment rate is found by dividing the estimated $\Delta ln(P_t)$ coefficients by 100. For example, for the period 1904–29 the coefficient of $\Delta ln(P_t)$ is −10.84 which is interpeted to mean that when the inflation rate goes from, say, 0 percent to 1 percent, the unemployment rate falls initially by about .11 percentage points.

from [1, p. 262] and [20]. Hourly compensation was obtained by linking the Lucas-Rapping [9] all economy hourly money compensation series (1929–65) to the Rees [20] series for manufacturing (1900–29).[13]

In Table 1 we report estimates of the coefficients of (8) for the period 1904–65 (line 1), for three subperiods (lines 2–4) and roughly comparable estimates for 1930–65 from [9] (line 5).[14] The division of the period into three subperiods is arbitrary. We wish to separate the postwar years to obtain comparability with the many Phillips curve studies done for this period. In view of the folklore referring to 1929 as "the end of an era," it seemed another natural dividing point. Whatever the reasons for this subdivision, however, we think the reader will agree after seeing the results that it was fortunate that we tried it.

In Table 1 standard errors are given below each coefficient estimate. Each set of estimates indicates a (short and long run) Phillips curve with a quantitatively significant negative slope. Thus for example, leaving aside the question of statistical significance, the estimated inflation term for the period 1946–65 implies an initial fall in the measured unemployment rate of about .14 percentage points per percentage point increase in the inflation rate. Under a sustained inflation, the final impact on unemployment is .18 per percentage point increase in the inflation rate [.1380/(1−.249)]. The real wage rate has a statistically significant effect in the predicted direction for the entire period and for the 1904–29 period, but not for the latter two periods. There is an indication of serial correlation in each regression.[15] The Chi-

[13] The data used in this paper are available upon request from the authors.

[14] In [9], we measured price by the GNP deflator.

[15] We refer to the stated values of the Durbin-Wat-

TABLE 2—ESTIMATES FOR MODEL BASED ON EQUATIONS (5), (6) AND (9).

Period	Constant	$\Delta ln(P_t)$	$\Delta ln(P_{t-1})$	$\Delta ln(P_{t-2})$	$\Delta ln(w_t)$	$\Delta ln(w_{t-1})$	U_{t-1}	U_{t-2}	R^2 (Adj)	DWS
1904–65	1.15	−39.16	35.36	−17.38	−11.86	26.67	1.170	−.322	.876	2.01
	(.63)	(6.69)***	(8.19)***	(6.84)***	(9.19)	(9.52)***	(.126)***	(.123)***		
1904–29	2.12	−20.09	30.32	−12.61	−34.78	− 1.32	.680	.021	.532	1.91
	(1.21)	(8.11)**	(9.07)***	(9.11)	(13.52)***	(12.92)	(.227)***	(.232)		
1930–45	1.98	−96.87	5.97	−24.27	18.90	11.44	.423	.374	.966	1.87
	(2.08)	(15.49)***	(22.13)	(13.04)	(22.24)	(21.50)	(.277)	(.286)		
1946–65	5.55	−12.73	−13.20	7.00	−15.05	−10.44	.247	−.147	−.05	1.94
	(3.29)	(17.29)	(34.20)	(15.48)	(22.68)	(30.39)	(.305)	(.275)		

** } one tail significance at { .01
*** } one tail significance at { .005

square statistic for testing the stability of the coefficients across the three subperiods takes the value 70.9, to be compared to the .005 critical value (8 degrees of freedom) of 22.0.[16] Hence the hypothesis that the coefficients are equal in all subperiods is decisively rejected.

In turning to estimates for the more complicated hypothesis (11) (with $k=2$, $n=1$, $m=2$) reported in Table 2, we have in mind three sets of questions. First, is the additional complexity justified by significance tests on the added coefficients? Second, are the instability and serial correlation which appear in the Table 1 estimates eliminated? Third, do the estimates succeed in reconciling a downward sloping short run Phillips curve with a flat long run curve?

The estimates for the entire period, and to a lesser extent for the subperiods, indicate clearly that the additional lagged inflation rates, rates of wage change, and unemployment rates result in a significant improvement.[17] Further, serial correlation appears to be absent from all regressions. The instability across periods is, however, still present as measured by the Chi-square statistic 64.2, to be compared to the .005 critical value (16 degrees of freedom) of 34.3. Thus, we must discuss the results for each period separately, rather than treat the first line of Table 2 as an accurate Phillips curve for the entire period. (It is fortunate that we chose arbitrarily to split the data period since there is no clue in the results on the first line of Table 2 that the regression over the whole period is dangerously misleading.)

son statistic. Of course, the critical values tabulated by these authors in [2] do not apply to stochastic difference equations as estimated here.

[16] Let SS_Ω and DF be the residual sum of squares and the degrees of freedom under the maintained hypothesis, let SS_ω be the residual sum of squares under the null hypothesis, and let df be the number of independent restrictions imposed by the null hypothesis. The statistic used is $(DF)(SS_\omega\text{-}SS_\Omega/SS_\Omega)$ which converges in distribution to a Chi-square with df degrees of freedom as DF goes to infinity. In the subperiods, DF is as low as 8 in some of our tests, so *caveat arbiter*.

As an alternative test statistic, we considered the F-statistic based on SS_Ω and SS_ω. Since the estimated equations are stochastic difference equations, the test is also at best approximately valid and the properties of the latter approximation, while perhaps superior to the one we used, are unknown. Similar considerations led us to use unit normal, rather than Student's tables, to judge the significance of the estimates reported in Tables 1 and 2.

[17] The significance of coefficients of added variables is the criterion we have used to select the "best" of the various instances of (11).

In the two subperiods, 1904–29 and 1930–45, indicated in Table 2 there is evidence of a short run Phillips curve: the coefficient of $\Delta ln(P_t)$ in each regression is negative and significantly different from zero. On the other hand, for the period 1946–65 there is no statistical evidence of a short run Phillips curve, but this regression does not improve on the comparable period results reported in Table 1. To test for the existence of a long run Phillips curve we consider three null hypotheses. First, the hypothesis that a sustained, constant percentage rate of inflation has no long run effect on unemployment is seen, from (11), to require that the coefficients of all price change variables (three in the regressions reported in Table 2) add to zero: we refer to this hypothesis (in Table 3, below) as H_0. The hypothesis that the coefficients of the lagged wage changes sum to zero is termed H_1. Finally, let H_2 be the hypothesis that both H_0 and H_1 are true.

TABLE 3—CHI-SQUARE TESTS

Period	Hypothesis Tested	H_0	H_1	H_2
1904–65	χ^2	8.360***	1.159	9.566**
1904–29	χ^2	.095	3.051	3.819
1930–45	χ^2	23.497***	.702	24.627***
1946–65	χ^2	.587	.279	.942

** / *** Null Hypothesis Rejected at .01 / .005 significance level.

Table 3 provides Chi-square statistics for testing these hypothesis, for the entire period and for each subperiod, within the version of (11) reported in Table 2. In view of the instability of coefficients across periods, only the subperiod results appear to be of interest. Table 3 shows that there is no evidence of a long run Phillips curve over the first subperiod 1904–29. In this period, the first-period impact of an increase in the inflation rate is a decrease in unemployment, but this initial effect is counteracted in the second year with an effect in the opposite direction. The third period effect is again negative. Adding the price change coefficients over the three terms yields a negative estimate in this case, but this sum does not differ significantly from zero.[18] Similar remarks hold for the wage change coefficients over the first two periods.

The results for the second period, 1930–45, indicate the presence of a long run Phillips curve. But the results for the 1946–65 period are strikingly different. For this period there is no evidence of a long run relationship, a not too surprising result in light of the fact that the individual coefficients are all statistically insignificant in this period.

III. *Conclusions*

The objective of this paper has been to articulate what we have called an expectations theory of the Phillips curve, and to develop and test some of its implications. Our primary emphasis has been on the possibility, strongly suggested by our theory, that the Phillips curve is a short run phenomenon, in the sense that a sus-

[18] To estimate the slope of the long run Phillips curve, one divides the sum of the price change coefficients by one minus the sum of the lagged unemployment coefficients. This estimate can be large even in cases where its numerator does not differ significantly from zero. Thus, for example, for the period 1904–29 this estimate is -7.96 which is smaller in absolute value than the initial effect of -20.09 (Table 2). On the other hand, for the periods 1930–45 and 1946–65 the long run point estimates are -567.34 and -21.03 respectively. These estimates are larger than the initial inflation impact as estimated by the leading inflation term in Table 2. Of course, to the extent that the lagged unemployment rates reflect the impact of forces other than lagged inflation rates, we may substantially overstate the long run effect of inflation on unemployment rates. A discussion of this and other estimation problems in models such as ours can be found in Griliches [5].

tained inflation will have a temporary but *not* a long run effect on unemployment. As reported in the last section, our results show (1) that our particular expectations-based Phillips curve is not stable over the three (arbitrary) subperiods of the period 1904–65, (2) that a short run Phillips curve exists in each subperiod, although the relationship for the most recent period, 1946–65, is less statistically obvious than for the earlier periods, and (3) that the hypothesis that there is no long run Phillips curve is accepted for the periods 1904–29 and 1946–65, but rejected for the period 1930–45.

What is to be concluded from these tests? As reported econometric models go, ours can scarcely be called successful, but we think its failures are suggestive along several lines. First, it is clear that the existence of a short run Phillips curve is consistent (in practice as well as in principle) with the absence of a long run inflation–unemployment tradeoff. Second, the expectations view as we have developed it appears to have promise in the sense that the additional explanatory variables it suggests are frequently (though by no means uniformly) significant. Third, and most important, statistical Phillips curves are highly unstable over time, and this instability is far too serious to be dismissed by a vacuous reference to structural change in either 1929 or 1946. Until the variables which are shifting these curves can be identified, and verified as important by testing over the entire period, we see no alternative to the conclusion that empirical Phillips curves (ours included) are a weak foundation on which to base policy decisions.

REFERENCES

[1] Council of Economic Advisors, *Economic Report of the President*, January 1967, Washington 1967.

[2] DURBIN, J. AND G. S. WATSON, "Testing for Serial Correlation in Least Squares Regression II," *Biometrika*, 1951, *38*, 159–77.

[3] ECKSTEIN, O. AND T. WILSON, "Determination of Money Wages in American Industry," *Quart. Jour. Econ.*, Aug. 1962, *76*, 379–414.

[4] FRIEDMAN, M., "The Role of Monetary Policy," *Am. Econ. Rev.*, March 1968, *58*, 1–17.

[5] GRILICHES, Z., "Distributed Lags: A Survey," *Econometrica*, Jan. 1967, *35*, 16–49.

[6] JORGENSON, D. W., "Rational Distributed Lag Functions," *Econometrica*, Jan. 1966, *34*, 135–149.

[7] LEBERGOTT, S., *Manpower in Economic Growth: The American Record Since 1800*, New York, 1964.

[8] LIPSEY, R. G., "The Relationship Between Unemployment and the Rate of Change of Money Wage Rates in the U.K., 1862–1957," *Economica*, Feb. 1960, *27*, 1–31.

[9] LUCAS, R. E., JR. AND L. A. RAPPING, "Real Wages, Employment and Inflation," *Jour. Pol. Econ.* (forthcoming).

[10] MELTZER, A. H., "Is Secular Inflation Likely in the U.S.," *Monetary Problems of the Early 1960's: Review and Approval* Atlanta, 1967.

[11] McGUIRE, T. W. AND L. A. RAPPING, "The Role of Market Variables and Key Bargains in the Manufacturing Wage Determination Process," *Jour. Pol. Econ.*, Oct. 1968, *78*, 1015–36.

[12] ———, "The Determination of Money Wages in American Industry: Comment," *Quart. Jour. Econ.*, Nov. 1967, *81*, 684–90.

[13] ———, "The Supply of Labor and Manufacturing Wage Determination in the United States: An Empirical Examination," *International Econ. Rev.* (forth coming).

[14] ———, "Interindustry Wage Change Dispersion and the 'Spillover' Hypothesis," *Am. Econ. Rev.*, June 1966, *61*, 493–501.

[15] MORTENSON, D. T. "A Theory of Wage and Employment Dynamics," in E. Phelps et. al., *The New Microeconomics*

in Employment and Inflation Theory: Price Wage and Employment Decisions, forthcoming.

[16] MUTH, J. F., "Rational Expectations and the Theory of Price Movements," *Econometrica*, 1961, *29*, 315–35.

[17] PERRY, G., "The Determinants of Wage Rate Changes," *Rev. Econ. Stud.*, Oct. 1964, *31*, 287–308.

[18] PHELPS, E. S., "Money Wage Dynamics and Labor Market Equilibrium," *Jour. Pol. Econ.*, Aug. 1968, *76*, Part II, 687–711.

[19] PHILLIPS, A. W., "The Relation Between Unemployment and the Rate of Change of Money Wage Rates in the United Kingdom, 1861–1957," *Economica*, Nov. 1958, *25*, 283–99.

[20] REES, A., "Patterns of Wages, Prices, and Productivity," in C. Myers (ed.) *Wages, Prices, Profits and Productivity*, New York 1959.

[21] TELLA, A. J., AND P. A. TINSLEY, "The Labor Market and Potential Output of the FRB-MIT Model: a Preliminary Report," (paper presented at the Winter, 1967 meeting of the Econometric Society).

COMMUNICATIONS

WAGES AND THE GUIDEPOSTS: COMMENT

By PAUL S. ANDERSON*

George Perry's first test of the impact of the wage guidepost policy of 1962–66 involved the prediction of wage rises in manufacturing for the 1961–66 period using his Phillips Curve regression equation based on the years 1947–60; he found that this equation overestimated 1962–66 wage rises by up to $2\frac{1}{2}$ percent on an annual basis [3, pp. 898–99]. Perry's results were corroborated in a test made by Levy, who found that Perry's Phillips Curve equation from 1948–67, with a guidepost dummy variable for the period 1962–67, yielded a coefficient of -1.5 for the dummy which was significant at .01 [1, p. 12].

Since these tests are based on Perry's version of the Phillips Curve regression equation, the reliability or stability of that equation is crucial to evaluating the significance of the test results. Shown in Table 1 are Phillips Curve regression equations with Perry's variables for the 1948–67 period and various subperiods which were derived by Levy [1, pp. 3, 12]. The first equation is essentially Perry's basic equation ending in 1960 while the next two extend to 1967 with a guidepost dummy in the third. The coefficients vary little among these three equations and are highly significant which suggests both that Perry's basic equation is stable and that the guidepost was effective.

However, when the 1948–67 period is divided into three subperiods, the coefficients become extremely unstable with two of them actually changing sign. It is not surprising that Levy found that they "failed" the Chow test that no statistically significant structural shift occurred between each successive pair of subperiods [1, p. 6]. In light of such instability there is serious doubt as to whether these independent variables bear any basic causal relation to wage changes. A more reasonable interpretation might be to consider all the variables, both dependent and independent, as business cycle indicators with the true independent variable being fluctuations in business conditions.

When a regression equation exhibits a marked structural shift, it is to be expected that a dummy variable for some part of the period spanned will turn out statistically significant. In order to test whether other periods than 1962–67 would yield a significant value for a dummy variable, the present author inserted the value unity for consecutive 5-year periods in turn over the entire 1948–67 span, beginning with each even-numbered year. The coefficients of the dummy were positive for the first part of the 1948–67 period and then turned zero and negative. Significant values for the dummy were obtained for 1956–60 (positive) and 1962–66 (negative). Since the dummy for the 1960–64 period was almost zero while that for 1962–67 was highly significant, the 1962–67 period was divided into two dummy test periods, 1962–64 and 1965–67. The dummy for 1962–64 was not significant (t value of -1.1) while that for 1965–67 was highly significant (t value of -7.1).

To summarize these various dummy tests: (1) all that a significant value for a dummy variable shows is that a significant structural shift has occurred and Levy's Chow test showed this had happened to Perry's Phillips Curve equation even before 1962; (2) a significant value was obtained for a dummy variable inserted into the 1948–67 equation for 1956–60; and (3) a dummy variable for 1962–64 was not significant. It seems reasonable to conclude that dummy variable testing leaves the efficacy of the wage guidepost policy in doubt.

Perry's next demonstration of the impact of the wage guidepost involved a comparison of wage rises in visible and invisible manufacturing industries, with visible industries being those judged most susceptible to guidepost pressure. Perry showed these industries had statistically significantly higher ratios of

* The author is financial economist at the Federal Reserve Bank of Boston.

TABLE 1—COEFFICIENTS OF PHILLIPS CURVE REGRESSION EQUATION

Period	Constant	Change in CPI	Profit Rate	Change in Profit Rate	Inverted Unempl. Rate	Guidepost Dummy†
1948–60	−4.7	.35**	.44**	.73**	15.0**	
1948–67	−2.8	.42**	.27**	.49*	13.7**	
1948–67	−2.6	.40**	.31**	.66**	12.4**	−1.46**
1948–53	−17.9	−.04	1.18**	1.01**	33.6**	
1953–60	−4.5	.63**	.31*	1.09**	19.6**	
1960–67	3.1	1.01**	−.30	.53	7.6	
1960–67	2.5	1.00**	−.18	.68*	5.0	−0.42

* Significant at .05.
** Significant at .01.
† Value of 1 for IIIQ 1962–IQ 1967.
Source: [1, pp. 3, 12].

wage rises in 1954–57 to wage rises in 1963–66 than did the invisible industries [3, p. 900].

Perry admitted his classification of industries by visibility was necessarily subjective because no objective measure was available. Two possible objective measures of visibility are the concentration ratio and the percentage of workers unionized. The concentration ratio might be the more relevant measure for visibility with regard to the price guidepost while unionization is probably more relevant to the wage guidepost. If Lewis' percentage of workers unionized [2, p. 274] had been used to classify manufacturing industries with 50 percent being the dividing line, several industries in Perry's list would have been reclassified—machinery, chemicals, and food to the invisible group and tobacco and apparel to the visible group. These shifts would have lowered the difference between the means of the ratios of wage rises in the two periods to near statistical insignificance (to around .10 level for the 1954–57 and 1963–66 ratio, for example).

Implied in Perry's visible-invisible test is the interesting assumption that wages in the visible industries would normally be expected to rise at a greater percentage rate than those in the invisible industries. Accordingly, the gap between wages in highly-unionized industries and those in predominantly unorganized industries should widen continuously, except perhaps when guideposts are used. As Lewis pointed out, the union-nonunion wage gap has not shown any discernible trend since 1919; at times it has narrowed and at other times it has widened [2, p. 207]. If the assumption about continual widening of the gap is replaced by the assumption that the size of the gap fluctuates about some normal level, then the difference between the 1954–57 and 1963–66 wage rises of the visible and invisible industries might be explained by the old regression fallacy argument: if variables are classified according to their departure from some normal in one time period, they are likely to show a shift toward normal in the next time period [4, pp. 258–63]. This argument can be applied to Perry's visible-invisible groups by observing how the ratios of the average wages of each group to the manufacturing average changed over time (basic data obtained from [7, pp. 148–49]).

	1954	1957	1963	1966
Visible group	1.117	1.122	1.137	1.125
Invisible group	.853	.843	.834	.843

If the 1957 and 1966 indexes were judged about normal, the 1954–57 and 1963–66 changes would be classified as movements from abnormal toward normal levels.

Perry's tests were all indirect or inferential. One approach to more direct testing is to analyze data covering actual changes in wage rates to see if they show any indication of the impact of the guidepost. Since the aim

TABLE 2—PERCENTAGE FREQUENCY DISTRIBUTION OF WAGE CHANGES IN MANUFACTURING
1959–66 Percentage Change

Year or Period	0	0.1–0.9	1.0–1.9	2.0–2.9	3.0–3.9	4.0–4.9	5.0–5.9	6.0–6.9	7.0–7.9	8.0–8.9	9.0–9.9	10.0+
						Union						
1959	13	11	3	10	32	17	8	3	2		1	
1960	13	1	6	14	34	21	6	2	1	1		2
1961	17	4	8	27	28	8	5	1		1		1
1959–61	14	5	6	17	31	15	6	2	1	1		1
1962–66	21	1	9	25	23	8	5	4	1		1	
1962	27	1	7	27	28	5	3	1		1		
1963	22	2	8	27	25	5	4	3	1		3	
1964	24	2	18	26	20	4	4	1		1		
1965	13	1	8	28	26	14	5	2	1		1	
1966	19	1	6	19	17	12	10	12	2	1		1
						Nonunion						
1959	31	1	3	8	18	11	12	6	5	2		1
1960	41	2	3	9	18	10	12	2	1	1		1
1961	46	4	4	11	17	6	4	1	1	1		6
1959–61	39	2	3	9	18	9	9	3	2	1		3
1962–66	34	2	4	14	17	11	11	3	2	1		2
1962	47	1	6	17	15	8	4	1				1
1963	30	1	3	17	17	8	12	4	2	3		2
1964	44	2	3	14	18	9	8	1	1			1
1965	25	1	4	15	18	14	15	3	2	1	1	3
1966*	22	1	3	8	15	16	18	6	4	1	1	2

* 4% not specified or not computed.
Source: [5] and [6].

of the guidepost policy was to persuade union-management wage negotiators to keep wage rises from exceeding 3.2 percent annually, evidence of success would be found in the behavior of wage-rate changes at, and above, 3.2 percent. The average, or mean change which Perry analyzed, is almost irrelevant. For example, if the mean change in the 1962–66 period was below some standard only because there was a greater frequency of no-increase settlements, while the frequency of changes above 3.2 percent remained unaffected, the guidepost policy could hardly be given the credit. And yet, if the frequency of zero settlements declined and the mean was lower only because the right (upper) tail of the distribution of wage changes was truncated or cut off, then the guidepost policy should be given a great deal of credit.

Frequency distributions of union and nonunion wage changes have been compiled by the Bureau of Labor Statistics since 1959 and are presented in Table 2. These distributions of wage changes are based on reports from manufacturing establishments employing 12 million production and related workers where general wage changes are made. Of these 12 million workers, 9.3 million are in establishments covered by union agreements. Excluded are 2.1 million nonunion workers in plants where wage changes

are made only on an individual basis. In the table, column headings show wage changes in percent, ranging from zero, or no change, to 10 percent and above. Thus, reading the first row, 13 percent of union workers in manufacturing received no wage rise in 1959, 11 percent received wage rises of less than one percent, etc. Averages for 1959–61 and 1962–66 are presented for both union and nonunion distributions to permit ready comparison of pre-guidepost experience with the results after guideposts were announced.

There is a rich variety of information in these distributions and they could be analyzed in several different ways, but the conclusions appear quite simple and straightforward. While there was a decline in the proportion of union settlements exceeding 4 percent from a 1959–61 average of 26 percent to a 1962–66 average of 19 percent (see rows 4 and 5), this decline reflected a leftward shift in the entire distribution. The zero settlement class rose from 14 to 21 percent and the 3.0–3.9 percent increase class, which should have risen if the guidepost reduced the tail above 3.2 percent, actually declined from 31 to 23 percent. The nonunion distributions verify the connection between the frequency of the zero settlement class and the frequency of the tail above 4 percent—whenever the zero class increased, the above-4 tail decreased, and vice versa. Thus the guidepost did not have a noticeable effect on the distribution of wage settlements, and in particular, there was no unusual bulge around 3.2 percent which would suggest that settlements which would otherwise have been made for increases above 3.2 percent had been reduced to 3.2 percent.

An interesting sidelight to the wage guidepost is, that while almost all discussion has centered on its efficacy as a ceiling, it could equally well be considered a floor. Some employer representatives have complained that governmental emphasis on the sufficiency and justice of a 3.2 percent annual wage rise tended to prod unions into rejecting lower offers. If the guidepost had been followed exactly, all wage changes would have been for a 3.2 percent increase. The actual mean of union wage settlements in manufacturing for 1962–66 as calculated from Table 2 was only 2.6 percent on an annual basis. Thus, so far as the anti-inflationary purpose of guideposts is concerned, the less than 100 percent adherence to the guidepost was a benefit on balance.[1]

REFERENCES

1. M. E. LEVY, "Full Employment Without Inflation," (mimeographed), address presented at Annual Meeting of Business Economists, Detroit Sept. 29, 1967. An abstract of this address may be found in the NICB *Record*, Nov. 1967, *4*, 36–41.
2. H. G. LEWIS, *Unionism and Relative Wages in the United States*, Chicago 1963.
3. G. L. PERRY, "Wages and the Guideposts," *Am. Econ. Rev.*, Sept. 1967, *57*, 897–904.
4. W. A. WALLIS AND H. V. ROBERTS, *Statistics: A New Approach*, Glencoe 1956.
5. *Mon. Lab. Rev.*, Dec. 1964, *87*, 1404.
6. ———., Aug. 1967, *90*, 34.
7. U. S. DEPARTMENT OF LABOR, Bureau of Labor Statistics, *Handbook of Labor Statistics 1967*.

[1] Actual hourly earnings in manufacturing rose at a 3.2 percent annual rate for 1961–66 [7, p. 147], but this reflects overtime earnings, shifts in manufacturing structure, and upgrading of workers within plants as well as rises in basic wage rates.

WAGES AND THE GUIDEPOSTS: COMMENT

By MICHAEL L. WACHTER*

In a recent article George Perry supports the efficacy of the guideposts policy [4]. The purpose of this note is to show that Perry's choice of tests is arbitrary and that other explanations of the wage behavior of the 1962–67 period do as well or better than the "guideposts hypothesis."

Perry draws his evidence from two sources. First, he uses a wage equation initially presented in [3], to predict the wage changes for the guideposts period. He finds that the

* The author is a PhD candidate at Harvard University. He wishes to thank J. T. Dunlop, E. Kuh, O. Eckstein, L. Thurow, and F. Ripley for their helpful suggestions.

residuals after the second quarter of 1962 are all negative, indicating that predicted wage changes are greater than actual wage changes. This line of evidence can be criticized on the grounds that it is not clear that negative residuals indicate the success of the guideposts policy. Furthermore, even if Perry's test is correct, it can be shown that the forecasting equation is unstable and, related to this, that the long run of negative residuals is a statistical artifact.

Second, Perry divides the two-digit industries into "visible" industries (that is, those that are sensitive to the guideposts) and "invisible" industries. He finds that, although the wages for almost all industries increased at a slower rate during 1963–66 than during 1954–57, the slowdown was greater for the visible industries. The major problem with this evidence is that the visible–invisible classification matches too closely a division based on industry wage levels. This note will discuss the second point first and then analyze the stability properties of Perry's forecasting equation.

I. *Visible-Invisible Argument*

The main thrust of Perry's visible-invisible hypothesis is that the guideposts had a discriminatory impact on the rate of growth of industry wages, slowing down the rate of increase of some, while having little or no restraining effect on others. The industries classified as visible, however, have the common characteristic, among others, of relatively high wage levels. Therefore, it is possible that Perry's results reflect changes in the industry wage structure that emanate from sources other than the guideposts.

It has been suggested that from World War II until the late 1950's the industrial wage structure readjusted to the compression that occurred during the war. In fact, the coefficient of variation of the structure increased from .1293 in 1947 to .1857 in 1960. It then slowly retreated to .1812 in 1966.[1] As a result, Perry's *RATIO* variable (the ratio of wage change in 1954–57 to wage change in 1963–66) reflects the fact that high wage industries were increasing their wages faster, during the 1950's, than the low wage industries and that this tendency did not continue into the 1960's.

To test this, I regressed the wage change ratio (*RATIO*), computed by Perry, on a dummy variable (*DUM*) that takes on the value one whenever the industry is visible and zero elsewhere, and on the 1956 wage level of the industry (W).

$$(1)\quad \underset{}{RATIO} = .969 + \underset{(.04)}{.003DUM} + \underset{(1.9)}{.023W} \qquad \bar{R}^2 = .45$$

When *DUM* is dropped from the equation, the adjusted R^2 improves.

$$(2)\quad RATIO = .963 + \underset{(4.1)}{.027W} \qquad \bar{R}^2 = .48$$

The above indicates that the higher the wage level in 1956, the greater the slowdown in wage change. That is, any hypothesis that predicts a narrowing of the wage structure in the 1963–66 period, relative to the 1954–57 period, will explain Perry's *RATIO* variable better than a guideposts dummy variable.[2]

[1] The coefficient of variation given in the text is not weighted by the size of the industry. It is a simple average of hourly earnings in the twenty-one, two-digit industries. The data source is [8].

Note that this argument assumes that average hourly earnings in the industries studied by Perry can be used as a proxy for wage rates. Evidence to support this position can be deduced from Throop's results [6]. Throop regresses earnings on a skill mix variable. The coefficient of the skill mix variable is greater than one. This indicates that the higher the average hourly earnings, the greater the wage paid for equal skill.

[2] See [6] for an explanation of the behavior of the industry wage structure utilizing unions and changes in skill mix among industries as the causal factors. In particular, the World War II compression of wages forced unions out of equilibrium with respect to the union-nonunion wage differential, with the result that the 1950's was a period of relatively large wage increases in the extensively unionized, high wage, visible industries. A second explanation of industry wage structure behavior, based on the competitive hypothesis, is found in [5].

To best explain these results in a guideposts context, the visible-invisible division may be discarded and reliance placed on the guideposts provision that ". . . in the interest of equity, wages of workers who are underpaid because of weak bargaining power (or other reasons) should rise faster than the average" [7, p. 121].

TABLE 1—SUBPERIOD REGRESSIONS

Period	Constant	U^{-1}	$\dot{C}_{-1}$	R_{-1}	ΔR	R^2	D.W.
1948–2	−15.8	31.1	.075	1.01	.800	.939	1.24
1953–2	(6.2)	(5.6)	(0.7)	(5.6)	(4.0		
1953–3	−8.63	16.9	.326	.742	.600	.956	2.41
1957–4	(8.4)	(9.4)	(3.7)	(9.1)	(2.2)		
1958–1	−3.29	56.4	.549	−.404	1.20	.869	1.37
1961–4	(2.8)	(4.8)	(6.1)	(2.1)	(7.4)		
1962–1	2.43	23.67	.566	−.564	.349	.731	1.51
1967–2	(1.9)	(1.3)	(1.4)	(1.6)	(0.6)		

Source: The data were provided by George de Menil, assistant professor of economics at Boston College. The data have been slightly revised since Perry's work.

II. *Stability of the Forecasting Equation*

In order to test the success of the guideposts policy, a model must be formulated to explain the mechanism through which the policy operates. Unfortunately, Perry is vague on this point. Implicit, however, in his choice of tests is the notion that the guideposts policy only shifts his forecasting equation. The equation is the following:

$$\dot{W}_t = a_0 + a_1 U_t^{-1} + a_2 \dot{C}_{t-1} + a_3 R_{t-1} + a_4 \Delta R_t \tag{3}$$

where $\dot{W}$ is the change in hourly earnings, U^{-1} is the reciprocal of the unemployment rate, $\dot{C}$ is the percentage change in the cost of living and R is the profit rate.[3] Equation (3) describes the wage determination process both before and after the initiation of the guideposts. The hypothesized impact of the policy is to shift the parameter values so that for given values of the explanatory variables, the post-1962 parameter structure would predict smaller wage changes than the pre-1962 structure.

The Council's phrasing of the policy suggests a different interpretation [7]. The goal of the policy is to render wage changes a function of the guideposts formula which, in turn, is based primarily on changes in aggregate productivity. As a result, if Perry's hypothesis (that the guideposts policy is successful) is correct, then the major independent variable in a wage equation, for the period 1962–67, should be changes in productivity. Thus in this interpretation of the manner in which the guideposts work, equation (3) misspecifies the wage setting mechanism of the post-1962 period, so that Perry's negative residuals do not affirm the guideposts success.

Assuming, however, that negative residuals are evidence of a successful guideposts policy, Perry's results are still open to criticism. The forecasting equation is unstable and, related to this, the "unusually" long run of negative residuals beginning in 1962, is sensitive to the choice of period used to estimate the forecasting equation.

A stable wage equation is important for Perry's analysis. If a stable wage equation exists for the postwar period and then first shifts around 1962, it is more suggestive of a guideposts impact than if the equation shifts every few years (including 1962).[4]

To test the stability of the equation, the period 1948–67 is divided into four subperiods. The results are given in Table 1.

It is clear that parameter values for individual variables vary substantially among subperiods. Covariance tests also confirm

[3] The particular method of constructing these variables and the lags imposed on the equation are described in [3].

[4] Perry discusses the possibility that his equation shifts after 1953–52 and after 1957–4 [3, pp. 70–79]. Various factors can be offered to explain, *ex post*, these shifts. See [1, pp. 139–41] for an explanation, without reference to the guideposts, of the shifts around 1962.

the fact that the parameter structure (all variables taken together) shifts among subperiods.[5]

It is interesting to note that none of the variables is significant in the last subperiod and the R^2 is at its lowest level. This could suggest that the guideposts policy has had a major impact on the wage setting mechanism. However, the statistical fit of the Perry equation has been declining since the second period and the profit variable first appears with the wrong sign in the third subperiod.[6] Furthermore a reason for the low T statistics for the explanatory variables in the last subperiod is an increase in multicollinearity.[7]

Perry's residual analysis, that a long string of negative residuals begin with the inception of the guideposts, can also be questioned. First, long runs of residuals with the same sign are not an unexpected result of an autoregressive scheme. (Note that the Durbin-Watson statistic for the period 1948–2 to 1960–3 is only 1.2 [4, p. 898].) Second, it is not a fair test to compare residuals from a forecasted period with residuals occurring during the estimated period. An example of a more proper test is to estimate the Perry equation for the period ending 1957–4 and forecast the last two subperiods. The results are given in Table 2.

The equation's forecasts result in sixteen consecutive positive residuals for the third subperiod (1958–61). For the guideposts subperiod we find nine more consecutive positive residuals and then thirteen negative ones. Thus, under a different specification of the time interval for estimation and for forecasting, the abnormally long run of consecutive negative residuals found by Perry, becomes shorter and is overshadowed by a longer string of positive residuals. Note that the residuals do not become negative until two years after the guideposts policy is initiated.

Table 2

Period	Residual	Period	Residual	Period	Residual
1958–1	1.34	1961–1	1.10	1964–1	0.48
2	1.93	2	2.33	2	−0.26
3	2.15	3	2.59	3	−0.13
4	2.34	4	2.59	4	−1.20
1959–1	2.30	1962–1	2.17	1965–1	−1.20
2	1.98	2	1.30	2	−1.12
3	0.70	3	0.51	3	−1.96
4	0.61	4	0.38	4	−1.57
1960–1	1.31	1963–1	0.86	1966–1	−2.48
2	0.79	2	0.77	2	−2.63
3	1.91	3	1.14	3	−2.56
4	1.80	4	0.90	4	−2.43
				1967–1	−1.50
				2	−1.00

III. *Summary*

The main point of this note is that Perry's evidence in support of the guideposts is ambiguous. First, his visible-invisible classification of industries conforms too closely to a high-low wage division to give supporting evidence of the impact of the guideposts. Second, the theory of how the guideposts operate, implicit in Perry's analysis, is questioned. Third, his forecasting equation is unstable prior to 1962 so that a further shift in 1962 is not prima facie evidence in support of the guideposts hypothesis. Finally, the analysis of the residuals indicates that negative values do not appear with the inception of the guideposts as claimed by Perry.

[5] The calculated values of the covariance tests, for the three hypothesized shifts are respectively 2.94, 11.32 and 7.37. The first is significant at the 95 percent level and the remaining two are significant at the 99 percent level. The data periods for the three tests are 1948–2 through 1957–4, 1953–3 through 1961–4 and 1958–1 through 1967–2.

[6] Another casualty of the subperiod analysis is the notion that the guideposts acted to shift the Phillips curve towards the origin. (The Phillips curve is derived from the Perry equation by setting ΔR and C_{-1} equal to constants and plotting W as a function of the intercept and U^{-1} for fixed values of R_{-1}). First, the negative sign for R_{-1} in the last two subperiods makes it impossible to calculate the location of the Phillips curve for those periods. Second, the large variation in the value of the constant term and the coefficient for U^{-1} render the Phillips curve highly unstable.

[7] The respective values, for the subperiod regressions, of the Chi-square statistic (derived from the determinant of the moment matrix) are 35.7, 14.5, 38.2 and 115.5. See [2] for an explanation of the test.

REFERENCES

1. O. ECKSTEIN, "Money Wage Determination Revisited," *Rev. Econ. Stud.*, Apr. 1968, *35*, 133–144.
2. D. E. FARRER ane R. R. GLAUBER, "Multicollonearity in Regression Analysis: The Problem Revisited," *Rev. Econ. Stat.*, Feb. 1967, *49*, 92–107.
3. G. L. PERRY, *Unemployment, Money Wage Rates and Inflation.* Cambridge 1966.
4. ———, "Wages and the Guideposts," *Am. Econ. Rev.*, Sept. 1967, *57*, 897–904.
5. M. W. REDER, "Wage Differentials: Theory and Measurement," *Aspects of Labor Economics, Universities—National Bureau Conference Series*, Princeton 1962.
6. A. W. THROOP, "The Union-Nonunion Wage Differential and Cost-Push Inflation," *Am. Econ. Rev.*, Mar. 1968, *58*, 79–99.
7. *Economic Report of the President, January 1967*, Washington 1962.
8. U.S. DEPARTMENT OF LABOR, Bureau of Labor Statistics, *Employment and Earnings for the United States, 1909–66*, Bull. No. 1312–5. Washington 1967.

WAGES AND THE GUIDEPOSTS: COMMENT

By Adrian W. Throop*

In a recent contribution to this *Review* [7], George Perry brings to light two pieces of evidence which, while admittedly not conclusive, seem to suggest that the guideposts have significantly slowed the pace of money wage increases. The purpose of this note is to provide an alternative interpretation of the evidence and to offer a more adequate test of the impact of the guideposts.

Perry found that after the guideposts were made known, the rate of change of manufacturing money wages fell consistently below what would have been predicted by a regression equation based on postwar quarterly data.[1] He showed moreover that after the guideposts were introduced wage changes were significantly less for industries apparently subject to guidepost pressures.

Perry consulted some experts to separate two-digit manufacturing industries into those in which settlements are highly visible, and therefore potentially subject to guidepost pressure, and others. Besides having settlements which are widely publicized, however, the visible industries also have a high proportion of union members and a large number of relatively skilled workers. These latter characteristics could have caused the wage behavior which Perry attributes to the guideposts.

As I have argued elsewhere [11], there are good reasons for believing that organized labor was in disequilibrium with respect to its relative wage at the beginning of the 1950's. The disequilibrium was caused by the rapid growth of union membership which took place in the early 1940's, during wartime wage controls, and by a relative rigidity of union wages during the large and unexpected demand inflation which followed. In the fifties, as unions were making the adjustment towards an equilibrium, nonunion wages advanced at a normal rate in relation to the level of unemployment, but union wages rose faster; therefore, the Phillips curve was shifted upward while this adjustment was being made.

The wage-push inflation would have ended

* The author is assistant professor of economics at Dartmouth College. He wishes to acknowledge his indebtedness to Colin D. Campbell, Martin Segal, and a referee for valuable comments.

[1] Otto Eckstein [3], using a model [4] which employs observations on wage rounds instead of quarterly observations, has obtained results which are similar to Perry's. Eckstein's model predicts a 6.5 percent annual rate of increase of money wages in durable goods industries for the 1963–66 wage round, compared with an actual rate of 2.7 percent. A partial explanation of the deflection of wage increases from the expected path has been offered by Simler and Tella [9]. They take account of demographic changes in the sixties which appear to have lowered reported unemployment relative to the amount of available labor reserves. An adjustment of the Perry and Eckstein equations to correct for the movement of unreported labor reserves reduces, but does not eliminate, the gap between predicted and actual wage values, however. There is still room for additional explanations running in terms of the guideposts and/or other hypotheses. For fuller treatment of this last point, see Eckstein [3] and Sheahan [8, Ch. 7].

once unions obtained an equilibrium relative wage for their members. Unions appear to have done so by 1958, on the average. However, the adjustment of unions was not a continuous one; in the fifties the union-nonunion wage differential widened mainly during years of high employment when the bargaining power of unions was the greatest [11, pp. 95–6]. A difference in the behavior of union wages would not, therefore, be observable until another period of high employment such as that which occurred subsequent to the announcement of the guideposts. On this interpretation, the relatively less inflationary movement of the wages of visible industries after the guideposts, compared with before, would be explained by the fact that these industries are highly unionized. And the over-prediction by Perry's regression equation of wage changes in manufacturing after the guideposts would be the natural consequence of a completion of the union adjustment.

There is another reason for expecting a weaker response of wages in visible industries to labor market conditions in the sixties than in the period before the guideposts. Skill differentials were no longer widening after the guideposts as they had been in the fifties.[2] This would have slowed wage changes in visible industries, whose workers are more highly skilled, relative to those in the earlier period.

In Section I a modified version of the model discussed in [11] is used to measure the behavior of the union-nonunion wage differential during the 1960's. This model is then employed in Section II as a framework for testing the impact of the guideposts on wages. Observations are made on Perry's grouping of visible and invisible industries as well as on an alternative grouping. Finally, Section III contains some concluding comments on guidepost policy.

I. *A Dynamic Model of Wage Structure*

The earlier model utilized an index of industry skill level to remove the influence of variations in skill mix and changes in occupational wage differentials from movements in average hourly earnings, before testing for the effect of unionism in the 1950's. This index of skill level was constructed for each census industry by weighting the current median annual earnings of various sex-occupational groups throughout the economy by their relative employments in the industry. Since data on industry employment by occupation come from the Census of Population, the change in this exact index during the 1960's obviously cannot be computed until after the 1970 Census. But a close approximation to it is possible.

The earlier index of industry skill level was defined for a census industry as

$$(1)\qquad S = \sum_{i=1}^{n} \frac{a_i}{E} w_i,$$

where

a_i = number of workers of the ith sex-occupational group employed by the industry,

E = total number of production or non-supervisory workers employed by the industry,

w_i = median annual earnings of workers in the ith sex-occupational classification throughout the economy.

In Table 1 the change in this index between 1950 and 1960 is decomposed for census industries in the previous sample according to the formula

$$(2)\qquad \Delta S = \sum_{i=1}^{n} w_i \Delta[a_i/E] + \sum_{i=1}^{n} \Delta w_i [a_i/E] + \sum_{i=1}^{n} \Delta w_i \Delta[a_i/E].$$

The first term is the amount of change which is directly due to changes in the industry's sex-occupational mix. The second is the change directly attributable to changes in median sex-occupational earnings. And the third consists of the interaction between these factors. The calculations demonstrate that changes in indexes of skill level in the 1950's were almost entirely dominated by

[2] For example, the ratio of the median annual earnings of male operatives to that of male laborers rose from 1.48 to 1.66 between 1950 and 1960, but remained at 1.65 in 1966. Similarly, the median annual earnings of male craftsmen relative to laborers climbed from 1.84 in 1950 to 2.11 in 1960, and then fell to 2.02 in 1966.

TABLE 1—COMPONENTS IN PERCENT OF THE CHANGE IN THE INDEX OF SKILL LEVEL: 1950 TO 1960

Industry — Changes due to:	Changes in the industry's sex-occupational mix[a]	Changes in median sex-occupational earnings[a]	Interaction between the two factors[a]
Wholesale Trade	1.1	98.0	.9
Retail Trade	−2.3	105.7	−3.4
Metal Mining	2.0	96.7	1.3
Bituminous Coal Mining	2.5	95.9	1.6
Crude Petroleum and Natural Gas	1.1	98.2	.7
Quarrying and Nonmetallic Mining	1.3	97.7	1.0
Contract Construction	.4	99.0	.6
Lumber and Wood Products	2.8	94.1	3.1
Furniture and Fixtures	− .4	100.4	0.0
Stone, Clay, and Glass Products	3.4	92.9	3.7
Primary Metals	2.3	95.1	2.6
Fabricated Metal Products	2.3	95.4	2.3
Machinery	.5	98.7	.8
Electrical Equipment	.8	98.3	.9
Transportation Equipment	2.2	95.7	2.1
Food and Kindred Products	2.2	95.1	2.7
Tobacco Manufactures	7.4	86.6	6.0
Textile Mill Products	.4	99.3	.3
Apparel and Related Products	−5.2	109.7	−4.5
Paper and Allied Products	2.7	95.3	2.0
Printing, Publishing, and Allied Industries	−1.1	101.7	− .6
Chemicals and Allied Products	4.4	91.2	4.4
Petroleum and Related Products	3.4	93.0	3.6
Rubber and Plastic Products	− .8	101.1	− .3
Leather and Leather Products	−3.7	107.2	−3.5

[a] For precise definition, see equation (2) in text.

variations in median sex-occupational earnings. Sex-occupational mixes were subject to only gradual change; the change in mix accounts for only 7 percent of the total change in the index even in the most extreme case. Thus, a good approximation to the change in an industry's index of skill level is obtained by simply weighting changes in median sex-occupational earnings by the industry's sex-occupational mix as of the nearest census date. This is the procedure which was adopted; we refer to this measure as the modified index of industry skill level.

The wage equation previously estimated for 1950–60 [11, Table 2] is[8]

$$\text{(3)}\quad \dot{W} = -.714 + \underset{(.281)^{**}}{1.85\dot{S}} + \underset{(.0549)^{*}}{.129U} + \underset{(.0292)^{**}}{.128LS}$$

$$\bar{R}^2 = .721^{**} \quad \bar{S}_e = .0590 \quad D.W. = 2.706^{**}$$

where

$\dot{W}$ = percentage change in average hourly earnings in a census industry,

$\dot{S}$ = percentage change in index of skill level for production or nonsupervisory workers in a census industry,

U = proportion of a census industry's production or nonsupervisory work-

[8] Numbers in parentheses are standard errors of the regression coefficients. The symbol* or ** indicates a significant t statistic at the 5 or 1 percent level of significance on the basis of a single-tailed test.

The symbol * or ** on an $\bar{R}^2$ denotes that it is significantly different from zero at the level of 5 or 1 percent by the F-ratio test. The same notation on a Durbin-Watson statistic denotes that the null hypothesis of residual independence cannot be rejected at a 5 or 2 percent level of significance with a two-tailed test.

ers that is covered by collective bargaining agreements,

LS = an index of labor supply which varies with the rate of change of employment and an estimated elasticity of labor supply.

The coefficient of U in this equation is the estimated increase in the union-nonunion wage differential during the fifties.[4] When an estimation of this model was attempted for 1962–66, a period after the guideposts, the coefficient of the percentage change in the modified index of industry skill level became insignificantly negative. This poor showing was due partly to the fact that there is little variance in either changes in earnings or changes in skill level during the 1960's. Probably of greater importance was the relatively high unemployment in the early sixties, which seems to have weakened the relationship between industry earnings and skill level enough to deprive the model of any explanatory power. To get around this difficulty, the period was lengthened to 1957–66, a span whose terminal years are both ones of high employment. It is possible to draw some significant inferences about the impact of the guideposts on wages from the results.

Using the modified index of industry skill level, the estimated parameters of the model for 1957–66 are[5]

$$(4)\quad \dot{W} = -.0725 + \underset{(.714)}{1.011\dot{S}} - \underset{(.0379)}{.0188U} + \underset{(.0345)}{.0500LS}$$

$$\bar{R}^2 = .234^* \quad \bar{S}_e = .0417 \quad D.W. = 1.715^{**}.$$

[4] Only a slightly better fit is obtained if allowance is made for the effects of a pattern among certain major settlements in the durable goods industries. This wage pattern tended to produce conformity in cents-per-hour changes in the union wage rates of affected industries. Among industries influenced by the pattern, the degree of unionization ought to be more closely related to cents-per-hour than to percentage changes in wages. In our earlier study, cents-per-hour changes in wages were regressed on the absolute change in the index of skill level, the degree of unionization, and the labor supply index multiplied by initial average earnings in an attempt to determine how strong the effects of the pattern were. It is apparent, however, that this procedure leads to a biased estimate of the coefficient of U due to a specification error. The relationship between W and S in our basic model is nonlinear. A nonlinear relation between ΔW and ΔS is implied also; and therefore; forcing the equation into linear form biases the estimated coefficient of U.

A more satisfactory approach is the following. In industries affected by the cents-per-hour wage pattern, the percentage change in average earnings due to unionism should be linearly related to U/W rather than to U. Since random forces automatically tend to produce a negative correlation between W and $\dot{W}$, however, an estimate of the initial wage should be substituted for W; this is derived from [11, Table 1]. Assuming that the average change in the union-nonunion wage differential within the pattern is the same as outside it, the appropriate unionization variable for industries outside the pattern is the degree of unionization divided by the average of estimated wages within the pattern. The resulting equation is

$$\dot{W} = -.726 + \underset{(.270)^{**}}{1.85\dot{S}} + \underset{(8.4)^{**}}{22.3U/\hat{W}} + \underset{(.0285)^{**}}{.130LS}$$

$$\bar{R}^2 = .737^* \quad \bar{S}_e = .0573 \quad D.W. = 2.65^{**},$$

where $U/\hat{W}$ is the proportion of workers that is covered by collective bargaining agreements divided by either a) estimated average earnings in the case of pattern affected industries, or b) the average of the estimated average earnings in the pattern in the case of other industries.

The coefficient of $U/\hat{W}$ is an estimate of the cents-per-hour wage increase due to unionism which was experienced within the pattern. A value of 22.3 cents is consistent with the previously estimated 12.9 percent increase in the union-nonunion wage differential. Nevertheless, the improvement in the corrected R^2 is minimal. While there is convincing evidence of an interindustry wage pattern among certain major settlements in durables during the fifties, diffusion to other bargains within each two digit classification seems to have been very limited indeed. The choice of industries assumed to have been affected by the pattern was based on Maher [6]. They are primary metals, fabricated metal products, machinery, electrical machinery, transportation equipment, food and kindred products, rubber and plastic products, petroleum and related products, and stone, clay, and glass products.

[5] The sample is the same as in [11] except that observations on contract construction and tobacco manufactures were omitted because of exceedingly large residuals. Published figures on straight time average hourly earnings had become available for the manufacturing industries in the sample. For others, use was made of BLS adjustment factors [13] for eliminating premium overtime from average hourly earnings. Sources of the employment and wage data are [16] and [17]. Median annual sex-occupational earnings for 1957 and 1966 were taken from [15], while 1960 employment weights for the modified index of industry skill level are from [14]. Otherwise sources and definitions are the same as in [11].

There is a suggestion of a small decline in the union-nonunion wage differential, a possibility which will be considered again. The corrected R^2 is considerably below that for the data of the 1950's, though still significant. As is well known, $\bar{R}^2$'s fall if the variance of the dependent variable is reduced while the distribution of the error term remains unchanged. This is the nature of the present situation. The standard deviation of the percentage change in average earnings is 10.6 for 1950–60, but only 4.6 in 1957–66. Still, the ratio of the standard error of the regression to the mean of the dependent variable is about the same, at .133 in 1957–66 versus .105 in 1950–60. Therefore, the estimates of the model's parameters for the new period are probably about as good as it is possible to obtain.

II. *A Test of the Impact of the Guideposts*

The wage equation estimated for 1957–66 permits us to predict what would have happened to the wages of various industries in the absence of the guideposts. A comparison of predicted with actual values should reveal the depressing influence of the guideposts on the wages of Perry's visible industries if there is any. A more powerful test statistically would be to include in the wage equation a dummy variable distinguishing between visible and invisible industries. This procedure would not allow many degrees of freedom if applied to Perry's grouping because Perry's sample is considerably smaller than ours. The significance of such a dummy variable for an alternative grouping is reported, below.

As shown in Table 2, only half of Perry's visible industries have negative residuals.[6] This should not be very surprising even to someone who believes that the guideposts did perhaps influence some settlements. Perry defines visible industries as those which are susceptible to the enforcement of the guideposts. Though all of Perry's visible industries may have been equally susceptible to guidepost pressure because of their prominence in the economy, Administration efforts to enforce the guideposts no doubt fell short of applying equal pressure to all. And as everyone knows, the Administration was much more successful in getting adherence to the guideposts in some settlements (e.g. steel and the merchant marine) than in others (e.g. automobiles, airline machinists, and the New York subways). Thus, to test for the effect of the guideposts, those industries and bargains which are highly visible and in which there also appears to have been at least a partially successful enforcement of the guideposts ought to be separated from the rest.

My colleague, Martin Segal, qualifies as an expert witness on these matters, having served on the senior staff of the Council of Economic Advisers following the announcement of the guideposts. Before the residuals from equation (4) were computed, he was asked to pick out the industries that he felt had actually been affected by the guideposts. He named only four out of twenty-three in the sample. They are primary metals, electrical equipment, petroleum and coal products, and rubber and plastic products; all are visible in Perry's sense. Negotiations in steel were pushed to a conclusion in the White House, and import competition may also have affected wages in primary metals. Segal points out that in the case of electrical equipment, and perhaps to a lesser degree in the other two industries, management seems to have been influenced by the philosophy of the guideposts. As it happens, all four of these industries have negative residuals, showing that wages rose less than would have been predicted by conditions of labor supply and the movement of occupational earnings. There is only about a 6 percent probability that all four of the residuals could have been negative by chance.[7]

Next, a dummy variable with a value of

[6] Exactly the same industries have negative residuals if the insignificant degree of unionization variable is dropped from equation (4).

[7] The calculation assumes that the signs of the residuals in Segal's group of affected industries have a binomial distribution with equal probabilities under the null hypothesis of randomness. This assumption holds only approximately, however, because the probabilities are not exactly equal and successive trials are not strictly independent in small samples.

TABLE 2—RESIDUALS FROM EQUATION (4)

	Negative Residual	Positive Residual
Perry		
Visible Group	Primary Metals Electrical Equipment Petroleum and Related Products Rubber and Plastic Products	Fabricated Metal Products Machinery Transporation Equipment Chemicals and Allied Products
Invisible Group	Lumber and Wood Products Furniture and Fixtures Leather and Leather Products	Stone, Clay, and Glass Products Textile Mill Products Paper and Allied Products
Segal		
Affected Group	Primary Metals Electrical Equipment Petroleum and Related Products Rubber and Plastic Products	None
Unaffected Group	Wholesale Trade Bituminous Coal Mining Crude Petroleum and Natural Gas Lumber and Wood Products Furniture and Fixtures Apparel and Related Products Leather and Leather Products	Retail Trade Metal Mining Quarrying and Nonmetallic Mining Stone, Clay, and Glass Products Fabricated Metal Products Machinery Transportation Equipment Food and Kindred Products Textile Mill Products Paper and Allied Products Printing, Publishing, and Allied Industries Chemicals and Allied Products

one for the four industries picked by Segal and zero for others was added to equation (4). As shown in equations (5) and (6), its coefficient turns out to be significantly negative at an acceptable confidence level. The guideposts are estimated to have reduced the rate of advance of money wages in affected industries by about 4 percent in 1957–66, or about 1 percent per year during the period of the guideposts themselves.[8] Interestingly enough, out of the four industries which are visible in Perry's sense but unaffected by the guideposts according to Segal, only one has a residual that is negative.[9] Visibility of the bargain seems to be a necessary but not a sufficient condition for wages to have been slowed by the guideposts. There is now a measured increase of about 1 percent in the union-nonunion wage differential although it is clearly not significantly different from zero; and in any case it does not at all compare with the 12 percent increase estimated for the 1950's. Thus, the

[8] Presumably, the guideposts primarily influenced the wages of union workers in each of the affected census industries. If so, a better dummy variable would be one taking on values equal to the degree of unionization (U) in the affected industries and zero otherwise. This dummy variable was in fact tested. Unfortunately, however, the values of U in the four affected industries are so similar that it is impossible to choose statistically between this dummy and the simpler one reported on in the text. Note, however, that the figure of 4 percent refers to the impact of the guideposts on average earnings in the affected census industries. The reduction in the rate of increase of union wages would be higher than that on the above assumption, at 4.8 percent.

[9] Exactly the same industries have negative residuals in equation (5) as in equation (6).

relatively moderate wage behavior of recent years appears to be connected with the lack of a continued growth in the differential between union and nonunion wages, something which is quite likely to have happened even without the guideposts.

$$(5)\quad \dot{W} = -.0726 + \underset{(.676)}{.983\dot{S}} + \underset{(.0392)}{.00986U} + \underset{(.0329)^*}{.0574LS} - \underset{(.0237)^*}{.0426D}$$

$$\bar{R}^2 = .315^{**} \quad \bar{S}_e = .0394 \quad D.W. = 1.627^{**}$$

$$(6)\quad \dot{W} = -.0649 + \underset{(.659)}{.978\dot{S}} + \underset{(.0230)^*}{.0543LS} - \underset{(.0212)^*}{.0403D}$$

$$\bar{R}^2 = .348^{**} \quad \bar{S}_e = .0384 \quad D.W. = 1.617^{**}$$

III. *Comments on Guidepost Policy*

It has been shown that the impact of the guideposts on wages may easily be misjudged because the natural conclusion of a temporary wage-push inflation was about due by the time the guideposts were announced. While the guideposts do seem to have affected some visible settlements, it is probably only coincidental that union wages, in general, ceased to grow relative to nonunion wages after the guideposts.

The completion of a temporary profit-push inflation raises a similar difficulty in evaluating the influence of the guideposts on prices. Leonard Weiss [12] recently demonstrated the existence of significant increases in the markups of concentrated industries from 1953 to 1959, with no tendency for either a reversal or further upward movement in them during 1958–63. He explains this markup inflation as a deferral of some of the earlier demand inflation.[10] Robert Solow [10] has confirmed the expected macroeconomic implication of Weiss' observations—a slower rate of increase of wholesale prices in the 1960's relative to changes in unit costs and the degree of capacity utilization than in the 1950's. Perhaps the guideposts should be given partial credit for preventing further increases in the markups of concentrated industries. On the other hand, the search of oligopolists for optimum prices had doubtlessly about run its course by the time of the guideposts.

The logic underlying the guideposts seems to be that in a high employment economy the monopoly power of labor and business generally tends to cause income claims to exceed the total value of output and so gives rise to a continuing inflationary spiral [1] [10]. The alternative view that income claims of monopolists may create a temporary, but not a permanent, inflationary bias at high employment is consistent with the evidence presented here. Since economic groups without market power are defenseless against a redistribution of income unfavorable to themselves, monopolists may always increase the relative price of their output if they think it is in their interest to do so. But the resulting cost-push inflation is basically a sectoral phenomenon which must eventually end of its own accord when monopolists reach an equilibrium. The irony of guidepost policy is the fact that postwar wage-push and profit-push pressures probably were already on the wane before the implementation of the guideposts.

[10] See also Adelman [2] and Friedman [5, pp. 228–29].

References

1. G. Ackley, "The Contribution of the Guideposts," in G. P. Shultz and R. Z. Aliber, eds., *Guidelines, Informal Controls, and the Market Place.* Chicago 1966.
2. M. A. Adelman, "Steel, Administered Prices and Inflation," *Quart. Jour. Econ.*, Feb. 1961, *75*, 16–40.
3. O. Eckstein, "Money Wage Determination Revisited," *Rev. Econ. Stud.*, April 1968, *35*, 133–43.
4. ——— and T. Wilson, "Determination of Money Wages in American Industry," *Quart. Jour. Econ.*, Aug. 1962, *86*, 379–414.
5. M. Friedman, "Some Comments on the Significance of Labor Unions for Economic Policy," in D. M. Wright, ed., *The Impact of the Union.* New York 1951.

6. J. E. MAHER, "The Wage Pattern in the United States, 1946–57", *Indus. Lab. Rel. Rev.*, Oct. 1961, *15*, 3–20.
7. G. L. PERRY, "Wages and the Guideposts," *Am. Econ. Rev.*, Sept. 1967, *57*, 897–904.
8. J. SHEAHAN, *The Wage-Price Guideposts.* Washington 1967.
9. N. J. SIMLER AND A. TELLA, "Labor Reserves and the Phillips Curve," *Rev. Econ. Stat.*, Feb. 1968, *50*, 32–49.
10. R. M. SOLOW, "The Case against the Case against the Guideposts," in G. P. Shultz and R. Z. Aliber, eds., *Guidelines, Informal Controls, and the Market Place.* Chicago 1966.
11. A. W. THROOP, "The Union-Nonunion Wage Differential and Cost-Push Inflation," *Am. Econ. Rev.*, Mar. 1968, *58*, 79–99.
12. L. W. WEISS, "Business Pricing Policies and Inflation Reconsidered," *Jour. Pol. Econ.*, Apr. 1966, *74*, 177–87.
13. "Eliminating Premium Overtime from Hourly Earnings in Manufacturing," *Mo. Lab. Rev.*, May 1950, *70*, 537–40.
14. U. S. BUREAU of the CENSUS, *Census of the Population, 1960, Special Report P-C (2)—7C on Occupation by Industry*, Tables 1 and 2. Washington 1963.
15. ———, *Current Population Reports: Consumer Income*, Ser. P-60, Nos. 30 and 50.
16. U. S. BUREAU of LABOR STATISTICS, *Employment and Earnings Statistics for the United States, 1909–66*, Bull. No. 1312-4. Washington 1966.
17. ———, *Mo. Lab. Rev.*, Apr., 1967, *90*, 87–100.

WAGES AND THE GUIDEPOSTS: REPLY

By GEORGE L. PERRY*

The comments in the three papers by Anderson, Throop and Wachter all, to some degree, criticize the evidence I presented on the effectiveness of guideposts [3]. All offer some evidence of their own which, with the partial exception of Throop's, aim to show the superiority of a nonguidepost explanation of recent wage behavior.

In the original paper, I went to some pains to describe my analysis as highly suggestive rather than conclusive and volunteered that some other explanation might be found to fit the facts better. Yet, after considering the criticisms in these papers and looking at two more years of wage behavior in light of the original analysis, I find myself about as persuaded as before that guideposts had a noticeable effect. The criticisms were not unanticipated; on close inspection, the new evidence offered does not seem to me to dominate the old; and more recent data seem to support further the original guidepost hypothesis.

Alternative Explanations of Relative Wages

Anderson notes that, in place of the industry breakdown chosen by my guidepost experts, a strict unionization breakdown would weaken the reported results. Of course it would, the more so the more accurate my classification is. I was aware of the existence of unionization (and industrial concentration) measures, but suspected the experts could make a more realistic division than some arbitrary weighting of such factors would provide. To be of any use, I (through the experts) had to divide the industries before looking at the data, and I never looked back. Now I am glad to see Anderson's findings confirm their choices.

Throop's paper presents the most elaborate model of an alternative hypothesis to guideposts, at the same time concluding there are some guidepost effects. He also notes that the visible industries are among the more unionized, and argues the mid-1950's were a period of catch-up for union wages. If so, and if this catch-up had been completed, a relative slowdown (higher wage-change ratio) would be expected for visible industries in the next high-employment period. But if this, rather than guideposts, were really the

* The author is professor of economics at the University of Minnesota. He is grateful to Edward M. Foster for his comments on this paper.

dominant influence, an industry-split by unionization would work better than the visible-invisible split, and Anderson tells us it does not. In addition, note that while Throop's hypothesis would predict a higher wage-change ratio as I have defined it, it would not predict slower growth in wages during the mid-1960's for visible compared to invisible industries. Yet wage increases were noticeably slower in the visible group.[1]

Throop's second argument is that the visible industries, on average, employ more highly skilled workers; that skill differentials in the wage structure, which had widened in the mid-1950's, had stopped widening and, indeed, were narrowing by the mid-1960's; and that this accounts for a good part of the wage behavior in visible industries without recourse to a guidepost explanation.

How much is lost here by circular reasoning? Throop clearly identifies the visibles as high skill industries. Using his equation to explain wage changes by industry, he finds a residual implying a guidepost effect in only 4 of the 9 visible industries once the narrowing skill differential in wages is accounted for. But he fails to consider whether guideposts helped narrow the differential itself through their effect on wages in the visible (skilled) industries. If guideposts had a differential effect and the correlation between skilled and affected industries were high enough, all the guidepost impact would appear in the variable Throop takes as independent and there would be no pattern at all to his residuals.

There are some other questions about the equation Throop uses to search for guidepost residuals. Of the three explanatory variables from his original equations fitted to 1950–60 data, the two for unionization and labor market tightness are unimportant in the 1957–66 equation he uses here. Something was going on to change things this sharply. Not only may guidepost effects be buried in the skill-wage variable as just noted, they may also get lost in changing the coefficient estimates for a period that includes them, particularly in an equation whose independent variables could be expected to be correlated with guidepost activity. Although it would still miss the possible impact of guideposts on the skill-wage variable, it might have been instructive to project wage changes in the guidepost period from the 1950–60 equation and see how forecast errors matched up with the visibility criterion. Throop does not report on this.

Wachter offers wage levels as an alternative to the visible-invisible categories for explaining relative wage movements. Wage levels are closely linked to skill-mix and unionization, and indeed, Wachter justifies their use by reference to Throop's earlier study [4]. So the preceding remarks apply here as well. But I must also comment on the specific statistical evidence Wachter presents. In a cross-section regression explaining my wage-change ratios by using industry wage levels and a dummy variable when an industry is classified as visible, Wachter reports the adjusted R^2 rises slightly when the dummy variable for visible industries is dropped. He does not report the result of dropping the wage-level variable instead. But, in any case, this kind of R^2 comparison is just not meaningful when a dummy variable is used as a proxy. The division into visibles and invisibles was a rough attempt to get at one dimension of the guidepost experience—the extent to which individual industries were directly affected by the government's concern. No doubt a correct list, if it were known to us, would be one weighted by *degree* of "visibility" rather than simply a two-way classification. When Wachter restricts only my variable and not his to a two-way classification, he may come out better. Who knows which would look better if they were both fully arithmetic indices?

If I restrict Wachter's wage level criteria to a two-way classification, it comes out worse than the two-way visible-invisible breakdown implied by the guidepost hypothesis. Two visible industries (rubber and plastics and electrical equipment) have wage levels (1966 data) below two invisible industries (paper and stone, clay, glass).

[1] [3, p. 901] where I note the average annual increase in visibles from 1963 to 1966 was 2.9 percent compared with 3.8 percent in invisibles.

Switching these, so that the visible group is changed to the nine industries with highest wage levels, weakens the explanatory power of the two-way classification.[2]

Finally, I would like to make a general comment on the methodology of all these arguments. With each of these attempts to substitute another explanation for guideposts—whether special forces associated with unionization, skill-wage differentials or wage levels—there exists a logical problem in identifying when the proposed special forces ceased to operate. They cannot be identified by looking at relative wage behavior in the 1960's, as is implicitly done, because that behavior may well have been the result of guideposts. Thus Anderson's table showing the ratio of visible and invisible industry wages to the all manufacturing average can more easily be viewed as a vivid illustration of what guideposts did than as an illustration of why a guidepost explanation is not needed. The same can be said of the coefficient of variation of the wage structure presented by Wachter and the observed behavior of the skill-wage differential noted by Throop. Nor is the intuition much guide as to when a particular wage will start marching in step with the average, witness the continuing surge in construction wages. As Throop points out, historically the observed catch-ups occurred during high employment periods, so that a respite after the mid-1950's would itself offer no evidence that an equilibrium had been reached.

This is not to say these alternative hypotheses could not be true, only that this kind of argument is more attractive the less it depends on actual wage developments in the 1960's to predict wage developments in the 1960's. Forecasting relative wage movements by using plausible a priori alternatives to the visible-invisible industry split gets at this indirectly. The evidence that has been presented is hardly decisive in favor of the alternatives. Finally, as I mentioned earlier, even if one had a firm hypothesis from other evidence that some sort of catch-up was completed in the mid-1950's, it would still not predict the slower growth of wages in visibles relative to invisibles that occurred in the 1960's.

The Distribution of Wage Changes

Anderson's statistics on the frequency distribution of wage changes bring a new kind of data to the question. Unfortunately, they are not "simple and straightforward" to interpret as he says. The increased percentage of workers with zero wage change in the 1962–66 period is the basis for his conclusion that guideposts were not the responsible factor in the wage moderation that occurred. But these data exclude all fringe benefits.[3] The Steelworkers, for example, got only fringes in both 1962 and 1963. They are in the zero column, but I would not conclude from that that guideposts were obviously not a factor in their settlement. On Anderson's interpretation, why is there no rise in the 0.1 to 1.9 percent increase range, but only in the zero column?

While we cannot distribute the value of fringes in the same way as wages, we can correct Anderson's interpretation of the bulge in the zero column. As one might guess, it turns out to reflect a shift to all fringe settlements by some unions. The big shift in Anderson's table occurs after 1961. Quoting from the U.S. Department of Labor, "... whereas in 1961 only one-fourth of those employed where wage rates were unchanged were affected by changes in supplementary benefits, the corresponding proportion in 1962 was almost one-half ..." [5, p. 4]. And again, with respect to 1963 labor market developments, "Where wages were left unchanged, liberalization in fringes affected 55 percent of the workers, ... " [6, p. 6]. The data referred to include nonunion establishments in which wages did not change, but the statistics are dominated by union contracts. For all workers employed where wage decisions were reached in the year

[2] The two being dropped have wage ratios of 1.018 and 1.201; the two being added both have 1.015 ratios. All data are from [3, p. 900] Table 2, using the ratios for three-year periods.

[3] Fringes are a problem for any analysis of wages. However, the problem seems greatest when trying to draw the kind of inference Anderson does about the zero column.

(which includes virtually all nonunion establishments) the proportion of decisions involving neither wage nor fringe increases can be calculated from available statistics. In contrast to Anderson's zero column showing a rise from 14 percent to 21 percent between 1959–61 and 1962–66, the proportion of workers receiving neither wage nor fringe adjustments falls from 14 percent in 1959–61 to 11 percent in 1962–66.[4]

With the zero column accounted for, the behavior of the crucial 2 percent to 5 percent increase columns in Anderson's table looks entirely consistent with a guidepost explanation. One cannot be sure, because there is no knowing how much fringes do to comparisons of those columns. But there is some margin of safety for my conjecture if one believes, as I do, that there are underlying determinants of wage change which would have normally produced a substantial rightward shift in the frequency distribution rather than the shift left that is observed.

I cannot agree with the assertions by Anderson and Wachter that the guideposts, if followed, would have put all wage settlements at 3.2 percent. In theory, the guideposts were concerned with the average rate of wage increase and explicitly did not disturb the distribution around the average [2, pp. 185–90]. In practice, the attempt was to moderate increases that threatened to be excessive by the standards of the theory as those standards could be applied in individual cases.

The Aggregate Wage Equation

I can only offer a few, rather imprecise, words in reply to the concern shown by Anderson and Wachter over the significance of the underpredictions in the aggregate wage equation. I have long been unhappy that the equation's residuals displayed considerable serial correlation and that the equation did not yield stable estimates over various short periods to which it might be fit. But I know of no better one. The relevant lag structure is no doubt more complex than in the equation. And the highly uneven incidence of major wage negotiations, together with the occasional existence of long-term settlements that make some wage changes a function of economic conditions considerably earlier, defy precise expression in an aggregate wage equation. Thus estimating over short periods will give predictably nonsense results. But over a sufficiently long period, we can hope to capture something like average estimates. From the nature of things, I have always preferred my 1947–60 equation over others fitted to shorter periods; and, in the paper being discussed, relied on the wage-change ratios covering three years in preference to those covering only a year or two.

When Anderson experiments with dummy variables by trying them for a large number of five year periods in the twenty-year interval 1948–67, he ignores the a priori aspect of the guidepost test. If a man told me the next five cards in a deck would be black and they were, I would think he knew something, even if I found a couple of other five card runs when I ran through the whole deck myself.

At that, Anderson finds a dummy significant for only one other period, 1956–60.[5] There probably is something unexpected about the late 1950's and early 1960's—I would guess a protracted effect from the inflation of the mid-1950's together with a possible tendency for positive residuals immediately after recessions, which show up both in 1958 and 1961. Wachter's forecasts from an equation fitted through 1957 are more disturbing in this respect. However, his equation is also subject to the criticism of *ex post* reasoning. He ends his regression period just when positive residuals are start-

[4] Data are from [1, p. 43] and Table 4 in annual issues of [5]. Anderson appears to have used the less relevant concept of "wage changes effective," which includes wage changes based on decisions reached in earlier years, rather than the "wage decisions reached" statistic used here. Because of this, and because the fringe data cannot be allocated to union and nonunion workers separately, the absolute percentages for different years are not directly comparable with Anderson's.

[5] The dummy for 1956–60 is positive. If a negative dummy really belonged in the guidepost period, omitting it would bias the rest of the period toward accepting a positive dummy variable, which raises further doubts on the significance of Anderson's manipulations.

ing in the longer period equation. I believe 1948 to 1957 is a bad period for a regression, being dominated by the rapid wage increases of the post war, Korean, and mid-1950 inflations and attributing too much of the wage increase to variation in the independent variables. I would have guessed this equation would under-predict wage changes consistently in a typical projection period. If so, the substantial drop in Wachter's forecast error, occurring in 1962, is more to the point than the fact the residuals remained slightly positive into 1964.

I should also clarify how I would read the results presented in the original paper since interpreting them is relevant to the criticisms and to some updated results presented below. The pattern of residuals from the aggregate wage-change equation would only gradually reflect the guideposts for two reasons: First, even if guideposts had a full effect from the start, they would show up gradually in the residuals from an aggregate wage equation as the fraction of all wages negotiated since their introduction rose from zero to one. If all wages were negotiated within a year, the full effect would be visible in the residuals after a year since the independent variable spans four quarters. Second, guidepost activities clearly were pursued with greater frequency and intensity during 1964–66 than during the preceding or following years. Together with the first reason for expecting a lag in the residual pattern, this leads us to expect *increasingly* negative residuals from 1964 to 1966; and with guideposts largely abandoned after that, *gradually* less-negative residuals thereafter. In evaluating the evidence of the residuals, these a priori expectations are important, though they are largely ignored in the criticisms that were made.

Recent Evidence: The End of Guideposts

My continued suspicion that guideposts affected recent wage behavior is supported by updating some of the results of the earlier paper. Table 1 brings the evidence of the residuals up to date. After 1966, when guideposts effectively expired, the residuals gradually return to insignificance, just as the guidepost hypothesis would predict. This behavior is not predicted by any alternative hypothesis.

TABLE 1—ACTUAL MINUS ESTIMATED PERCENTAGE WAGE RATE CHANGES

Year-Quarter	From 1947–60 Equation[a]
1962–1	0.84
–2	0.07
–3	−0.52
–4	−0.71
1963–1	−0.97
–2	−0.37
–3	−0.19
–4	−0.18
1964–1	−0.27
–2	−0.77
–3	−0.73
–4	−1.72
1965–1	−1.68
–2	−1.63
–3	−2.11
–4	−1.61
1966–1	−2.36
–2	−2.75
–3	−2.63
–4	−2.39
1967–1	−1.55
–2	−0.94
–3	−0.51
–4	−0.59
1968–1	+0.22
–2	+0.53
–3	−0.18

[a] $$\dot{W}_t = -4.313 + \underset{(0.054)}{0.367\dot{C}_{t-1}} + \underset{(2.188)}{14.711U_t^{-1}} + \underset{(0.068)}{0.424R_{t-1}} + \underset{(0.176)}{0.792\Delta R_t}, \qquad R^2 = .88$$

$\dot{W}$ is the percentage change in straight time hourly earnings over the past year in manufacturing

$\dot{C}$ is the percentage change in the Consumer Price Index over the year

U^{-1} is the reciprocal of the percentage unemployment rate over the year

R is the average profit rate in manufacturing over the year (after tax profits as a percentage of equity)

ΔR is the quarterly first difference in R

Standard errors of coefficient estimates appear in parenthesis.

REFERENCES

1. W. DAVIS AND L. M. DAVID, "Pattern of Wage and Benefit Changes in Manu-

facturing," *Mo. Lab. Rev.*, Feb. 1968, *91*, 40–48.
2. *Economic Report of the President, January 1962*, Washington 1962.
3. G. L. Perry, "Wages and the Guideposts," *Am. Econ. Rev.*, Sept. 1967, *57*, 897–904.
4. A. W. Throop, "The Union-Nonunion Wage Differential and Cost-Push Inflation," *Am. Econ. Rev.*, Mar. 1968, *58*, 79–99.
5. U. S. Department of Labor *Wage Developments in Manufacturing*, 1962 (Summary Release).
6. U. S. Department of Labor, *Wage Developments in Manufacturing*, 1963 (Summary Release).

TAX POLICY AND INVESTMENT BEHAVIOR: COMMENT

By Robert M. Coen*

In their recent study of tax incentives for investment, Robert E. Hall and Dale W. Jorgenson [6] report results which indicate that accelerated depreciation and the investment tax credit have been quite effective in stimulating capital expenditures in manufacturing. According to their model, the demand for capital depends, among other things, on an implicit rental price of capital which is partially determined by tax policy. Tax incentives influence the demand for capital solely through their effects on its rental price Hence, the crucial parameter in assessing the effects of tax policy is the price elasticity of the demand for capital. But, instead of presenting empirical estimates of this parameter, Hall and Jorgenson have imposed a unitary price elasticity by employing a Cobb-Douglas production function in their derivation of the demand function for capital. Part I of this note shows, using their own development of the theory and nearly identical data, that the price elasticity of demand estimated from the data is considerably below unity and that their estimates of the effects of tax policy are therefore much too high.

A second and perhaps more fundamental shortcoming of the Hall-Jorgenson study is that their procedure for calculating tax policy effects is not consistent with the theoretical development of their model. The model is based on the notion that a profit maximizing firm follows a special iterative decision process in which output happens to be a predetermined variable for the capital stock decision in any given period. This legitimizes their inclusion of output for estimation purposes as a determinant of the demand for capital, although it does not seem to me to be a rationalization of the acceleration principle, as Hall and Jorgenson [6, p. 392] and Jorgenson [7, pp. 40–41] have contended. While output may be predetermined for the capital stock decision, it is not an exogenous variable for a profit maximizing firm. Like labor input and capital stock, output is a decision variable, and its profit maximizing value depends on relative prices, including the implicit rental price of capital. In calculating the effects of tax policy, Hall and Jorgenson have, however, treated output as though it were an exogenous variable and ignored the effects of policy-induced changes in the rental price of capital on the firm's rate of output, thus violating the logic of their model.

This criticism should not be confused with another seemingly similar criticism that is applicable to all econometric studies of the effects of tax incentives for investment except that of Klein and Taubman [11], namely, that a single equation model cannot take account of the multiplier effects of investment generated by tax incentives. Higher investment means higher output, and higher output will in turn generate more investment. The reciprocal relation between investment and output cannot be ignored at the economy wide level, nor even at the level of certain industry aggregates, such as manufacturing; but it can generally be ignored at

* The author is a member of the faculty at Stanford University and is grateful to Charles R. Blitzer for computational assistance and to Bert G. Hickman and Robert Eisner for helpful comments on earlier drafts of this note.

the level of the firm. What cannot be ignored at the level of the firm is the fact that a fall in the price of a factor of production will, *ceteris paribus*, usually increase a firm's profit maximizing level of output. It is this type of output response that Hall and Jorgenson have overlooked.

Part II of this note elucidates the Hall-Jorgenson model and indicates the correct way of calculating effects of tax policy in that model. The magnitude of the errors committed by Hall and Jorgenson in calculating these effects cannot be assessed, however, without further extensive work with their model. It may be that the errors are small, but I will have to leave the burden of proof with them.

I. *The Elasticity of Substitution and the Effects of Tax Policy*

As Jorgenson [7], [8], [9] has shown, a competitive firm wishing to maximize its net worth should equate the marginal product of capital to the real rental price of capital, c/p, where c is the implicit rental price of capital (what Jorgenson originally referred to as the "user cost of capital") and p is the price of output. Much of the novelty of Hall and Jorgenson's work centers on the definition of c, which is shown to depend on tax policy.[1] They assume that the production function is Cobb–Douglas with an elasticity of output with respect to capital equal to α. Thus the marginal product of capital can be expressed as $\alpha Q/K$, where Q is output and K is capital stock; and upon equating this to the real rental price of capital, they obtain the expression for the desired level of capital stock

$$K^* = \alpha(p/c)Q. \tag{1}$$

It is well known that the elasticity of substitution in the Cobb–Douglas relation is equal to unity. Since the substitutability of capital and labor is an important determinant of the sensitivity of the demand for capital to changes in the rental price of capital, it hardly seems reasonable arbitrarily to assign the elasticity a value of unity—I say arbitrarily because no justification is given in their paper.[2]

An alternative to the Cobb-Douglas specification is the constant-elasticity-of-substitution (CES) production function,[3] which for two factors of production, capital and labor, constant returns to scale, and suitable choice of a unit of measurement for output can be written as

$$Q = [\beta K^{(\sigma-1)/\sigma} + (1-\beta)L^{(\sigma-1)/\sigma}]^{\sigma/(\sigma-1)}, \tag{2}$$

where L is labor input, β is a constant, and σ is the elasticity of substitution. Differentiating (2) partially with respect to K, equating the derivative to c/p, and solving for K, we have

$$K^* = \beta^\sigma (p/c)^\sigma Q. \tag{3}$$

Thus, the demand function for capital derived from the CES relation is of the same form as that derived from the Cobb-Douglas relation, except that the elasticity of substitution enters as an exponent on the relative price term. If the elasticity were zero, relative prices would not appear in (3); and since tax policy influences investment through

[1] The rental price measure adopted by Hall and Jorgenson is not the same as that used by Jorgenson [7] [8] and Jorgenson and Stephenson [10]. Jorgenson's original expression for the rental price [8, p. 249] was not an appropriate one for studying depreciation policy in the United States. It was correct only for the special case in which (a) the depreciation formula is of the declining balance form with a depreciation rate equal to the true economic depreciation rate and (b) policy makers stipulate that the depreciable base of acquisitions is a proportion v of their cost. A change in v would not change the timing of depreciation charges, but rather the total amount of the write off. This limitation of Jorgenson's formulation and the rental price measures relevant to U. S. experience were set forth in a paper entitled "Accelerated Depreciation, The Investment Tax Credit, and Investment Decisions," which I delivered at the December 1965 meetings of the Econometric Society. The rental price measure employed by Hall and Jorgenson corresponds to the one I have recommended.

There is an error in their formula for the present value of declining balance depreciation. The undepreciated balance at the point τ^+ is $e^{-(2/\tau)\tau^+}$, not $1-e^{-(2/\tau)\tau^+}$.

[2] Hall and Jorgenson actually assume a three factor Cobb-Douglas relation, which means that they must justify the assumption that the elasticity of substitution between any two of the factors—labor, equipment, and structures—is unity.

[3] This function was first derived by Arrow, Chenery, Minhas, and Solow [1] and by Brown and de Cani [2].

$$Q = [\beta_L L^{(\sigma-1)/\sigma} + \beta_B K_B^{(\sigma-1)/\sigma} + \beta_S K_S^{(\sigma-1)/\sigma}]^{\sigma/(\sigma-1)}, \tag{4}$$

its effects on c, there would be no investment response to changes in tax policy, according to the Hall-Jorgenson argument. If the elasticity is assumed to be unity, the results obtained can scarcely be regarded as an empirical test of the effectiveness of tax policy. The elasticity should be estimated, or compelling reasons should be given for assigning it a value of unity (or some other value). Evidence on the value of the elasticity is certainly very mixed; after an exhaustive examination of the evidence on the elasticity of substitution between capital and labor, Nerlove concluded that, "The major finding of this survey is the diversity of the results: Even slight variations in the period or concepts tend to produce drastically different estimates of the elasticity" [12, p. 58].

One way of determining the value of σ is to reestimate the Hall-Jorgenson model using (3) instead of (1) as the demand function for capital. Two problems are encountered, however. First, the least-squares estimating equations will be nonlinear in σ. Rather than carry out the difficult task of nonlinear estimation, I adopted a screening process to determine the value of σ; that is to say, I prespecified alternative values of σ, estimated the investment functions for each value, and examined the estimated equations to determine which value gave the best fit. The second problem is that there is no obvious generalization of the *CES* function for more than two factors. Hall and Jorgenson estimated separate investment relations for equipment and structures, even though they did not explicitly derive separate demand functions for each factor. It so happens that with the Cobb-Douglas assumption, the demand functions are related to (1) in a simple way; α should be replaced by the elasticity of output with respect to equipment (structures), c should be replaced by the rental price of equipment (structures), and everything else remains the same.

I have adopted a convenient but highly restrictive generalization of the *CES* function, namely (4).

where K_B is the stock of equipment, K_S is the stock of structures, and the β's are constants. According to this formulation, the elasticity of substitution between any two factors is the same and is equal to σ.[4] The demand functions derived from this form of the *CES* function are

$$K_B^* = \beta_B^{\sigma}(p/c_B)^{\sigma} Q \tag{5}$$

and

$$K_S^* = \beta_S^{\sigma}(p/c_S)^{\sigma} Q, \tag{6}$$

where c_B and c_S are, respectively, the rental price of equipment and the rental price of structures. In addition to estimating investment relations for equipment and structures separately, I have estimated a relation for total investment based on the demand function (3) with $c=(1.5c_B+c_S)/2.5$. The weights used in this expression for c are those suggested by the capital stock data in the Hall-Jorgenson paper.

The estimated relations, which are of the form

$$N_t = a_1\Delta K_t^* + a_2\Delta K_{t-1}^* + a_3 N_{t-1} + e_t, \tag{7}$$

where N is net investment expenditures, $\Delta K_t^* = K_t^* - K_{t-1}^*$, and e is a random disturbance, are presented in Table 1, along with the relevant Hall-Jorgenson estimates for comparison. The trial values of σ ranged from 0 to 2 in increments of 0.2.[5] The data I have used differ in two respects from the Hall-Jorgenson data: First, I used real gross product originating in manufacturing as an estimate of Q, and the implicit price deflator

[4] This generalization of the *CES* function has been suggested by Uzawa [14]. A less restrictive, but in some cases empirically manageable generalization is the so-called two-level *CES* function proposed by Sato [13]. If lagged adjustment is present, however, the strong separability of Sato's function—which so greatly simplifies its estimation—is lost, unless the adjustment rate for each factor is the same.

[5] Table 1 presents results only for values of σ between 0 and 1.2. For values of σ above 1.2, the uncorrected R^2 for the equipment regression fell steadily to .006 when σ was 2; that for structures remained negative; and that for equipment and structures combined fell steadily to .105.

TABLE 1—NET INVESTMENT REGRESSIONS $N_t = a_1 \Delta K_t^* + a_2 \Delta K_{t-1}^* + a_3 N_{t-1} + e_t$ WHERE $K_t^* = \beta^\sigma (p_t/c_t)^\sigma Q_t$

Annual data: 1951–63

A. Equipment

σ	0.0	0.2	0.4	0.6	0.8	1.0	1.2	Hall-Jorgenson[a] 1.0
$\hat{a}_1$	.0732	.0522	.0358	.0240	.0158	.0105	.0069	.0142
	(.0259)	(.0192)	(.0144)	(.0108)	(.0081)	(.0061)	(.0045)	(.0037)
$\hat{a}_2$	.0704	.0543	.0403	.0290	.0206	.0144	.0101	.0124
	(.0266)	(.0202)	(.0154)	(.0118)	(.0089)	(.0067)	(.0049)	(.0044)
$\hat{a}_3$	.6413	.6162	.6035	.6003	.6037	.6108	.6197	.6152
	(.1309)	(.1325)	(.1365)	(.1416)	(.1467)	(.1512)	(.1550)	(.1001)
R^2	.331	.343	.322	.282	.233	.184	.137	.722
$\bar{R}^2$	.198	.212	.187	.138	.080	.020	.000	—

B. Structures

σ	0.0	0.2	0.4	0.6	0.8	1.0	1.2	Hall-Jorgenson[a] 1.0
$\hat{a}_1$	.0384	.0275	.0175	.0099	.0052	.0026	.0012	.0040
	(.0127)	(.0087)	(.0062)	(.0044)	(.0031)	(.0021)	(.0014)	(.0013)
$\hat{a}_2$	.0308	.0260	.0192	.0130	.0085	.0054	.0035	.0053
	(.0127)	(.0084)	(.0058)	(.0040)	(.0028)	(.0019)	(.0013)	(.0014)
$\hat{a}_3$	.7537	.7201	.7121	.7253	.7483	.7723	.7937	.7658
	(.1076)	(.1017)	(.1019)	(.1065)	(.1117)	(.1156)	(.1181)	(.0790)
R^2	.138	.258	.263	.183	.077	Negative	Negative	.848
$\bar{R}^2$	.000	.110	.116	.019	.000	—	—	—

C. Equipment and Structures

σ	0.0	0.2	0.4	0.6	0.8	1.0	1.2
$\hat{a}_1$	.1093	.0769	.0512	.0327	.0203	.0123	.0074
	(.0306)	(.0227)	(.0171)	(.0130)	(.0097)	(.0072)	(.0053)
$\hat{a}_2$	.0960	.0756	.0568	.0414	.0296	.0210	.0148
	(.0313)	(.0233)	(.0176)	(.0134)	(.0100)	(.0075)	(.0055)
$\hat{a}_3$	.7173	.6896	.6743	.6697	.6727	.6801	.6896
	(.0989)	(.1004)	(.1048)	(.1107)	(.1165)	(.1215)	(.1255)
R^2	.475	.484	.456	.404	.346	.290	.240
$\bar{R}^2$	.370	.381	.347	.285	.215	.148	.087

[a] Annual data: 1931–41 and 1950–63

for gross product originating as an estimate of p,[6] whereas they used value added at factor cost as an estimate of pQ. The change of variables was necessitated by the fact that to fit the equations derived from the *CES* function, I had to split off the p from the Q, and I did not have a deflator for value added. Second, I did not have data on gross product originating for the prewar years and therefore used only postwar observations (1951–63), whereas Hall and Jorgenson used both prewar and postwar observations (1931–41 and 1950–63).

Several properties of the estimates should be noted. First, the best fits are obtained for values of the elasticity of substitution considerably below unity, 0.2 for equipment, 0.4 for structures, and 0.2 for equipment and structures.[7] Second, the parameter estimates obtained when σ is unity are very close to those reported by Hall and Jorgenson, which indicates that my use of a shorter sample period does not make much difference, that is, the relation proposed by Hall and Jorgenson does seem to be a stable one. Finally, the fits obtained to the postwar data are much poorer than those reported by Hall and Jorgenson for the prewar and postwar data.[8] Indeed, the poor explanation of post-

[6] I want to thank Gary Fromm and the Brookings Econometric Model Project for supplying me with these data.

[7] Because of the discrete nature of the trial values, these are not precise estimates of σ. Inspection of the R^2s in Table 1 suggests that best fits might be obtained for values of σ less than .2 for equipment and for equipment and structures and less than .4 for structures.

The results reported above corroborate those of Eisner and Nadiri [5] who have tested the basic Jorgenson investment model with quarterly data differing in several respects from that used here. Their tests also differ from my own in that they estimated the elasticity of K^* with respect to Q, while I constrained this elasticity to be unity, and they experimented with alternative lag distributions. Their best fitting regressions for manufacturing plant and equipment expenditures gave estimated elasticities of K^* with respect to p/c and Q equal to .1576 and .8158 respectively [5, Table 2].

[8] Since these regression lines are forced to pass through the origin, it is possible for the uncorrected R^2 to be negative, as in the structures regressions reported in Table 1. What this means is that over the sample period the variance of the dependent variable about its mean is smaller than its variance around the regression line, which would seem to indicate that the omission

TABLE 2—INVESTMENT ATTRIBUTABLE TO TAX INCENTIVES
(billions of 1954 dollars)

A. Equipment

Year	Hall-Jorgenson $\sigma=1.0$	Coen $\sigma=1.0$	Coen $\sigma=0.2$
1954	.418	.332	.125
1955	.680	.701	.218
1956	.480	.493	.146
1957	.305	.312	.090
1958	.154	.162	.045
1959	.124	.105	.032
1960	.133	.137	.039
1961	.089	.100	.064
1962	.821	.665	.276
1963	1.258	1.290	.393
Total	4.462	4.297	1.428

B. Structures

Year	Hall-Jorgenson $\sigma=1.0$	Coen $\sigma=1.0$	Coen $\sigma=0.4$
1954	.189	.133	.135
1955	.434	.404	.264
1956	.367	.353	.203
1957	.258	.248	.138
1958	.186	.185	.089
1959	.169	.160	.074
1960	.190	.186	.080
1961	.154	.159	.061
1962	.127	.130	.056
1963	.125	.127	.063
Total	2.199	2.085	1.163

C. Equipment and Structures

Year	Hall-Jorgenson[a] $\sigma=1.0$	Coen[a] $\sigma=1.0$	Coen[a] $\sigma=0.2$ for equip. $\sigma=0.4$ for struct.	Coen[b] $\sigma=1.0$	Coen[b] $\sigma=0.2$
1954	.607	.465	.260	.460	.194
1955	1.114	1.105	.482	1.165	.344
1956	.847	.846	.349	.900	.254
1957	.563	.560	.228	.605	.174
1958	.340	.347	.134	.379	.105
1959	.293	.265	.106	.274	.081
1960	.323	.323	.119	.324	.084
1961	.243	.259	.125	.251	.061
1962	.948	.795	.332	.733	.253
1963	1.383	1.417	.456	1.511	.403
Total	6.661	6.382	2.591	6.602	1.953

[a] Sum of separate estimates
[b] Overall estimates

$$L^* = A^{1/(1-\alpha-\beta)}[\beta(p/w)]^{(1-\alpha)/(1-\alpha-\beta)}[\alpha(p/c)]^{\alpha/(1-\alpha-\beta)}, \tag{13}$$

$$K^* = A^{1/(1-\alpha-\beta)}[\beta(p/w)]^{\beta/(1-\alpha-\beta)}[\alpha(p/c)]^{(1-\beta)/(1-\alpha-\beta)}, \tag{14}$$

war investment suggests that the model is faulty or incomplete.[9]

Table 2 presents estimates of investment attributable to tax incentives from the Hall-Jorgenson relations, from my relations assuming that σ equals one, and from my relations assuming that σ is equal to the value that gives the best fit to the data. The figures are obtained in the following way: Let $\tilde{c}$ be the rental price of capital that would have prevailed in the absence of tax policy changes. If tax policy had not changed, the desired capital stock would have been

$$\tilde{K}_t^* = \beta^{\sigma}(p_t/\tilde{c}_t)^{\sigma}Q_t, \tag{8}$$

and the change in investment resulting from tax policy changes is

$$\begin{aligned} N_t - \tilde{N}_t &= \hat{a}_1(\Delta K_t^* - \Delta\tilde{K}_t^*) \\ &+ \hat{a}_2(\Delta K_{t-1}^* - \Delta\tilde{K}_{t-1}^*) \\ &+ \hat{a}_3(N_{t-1} - \tilde{N}_{t-1}). \end{aligned} \tag{9}$$

This is not the correct way of calculating the effects of tax policy in the Hall-Jorgenson model, as I will show in Part II. It is, however, the procedure which they followed, and I want my results at this stage to be comparable with their estimates.

On the basis of the relations that best fit the data (and that do not arbitrarily assume that σ is one), one can hardly conclude that "tax policy is highly effective in changing the level and timing of investment expenditures" [6, p. 392]. Hall and Jorgenson estimate that tax incentives increased net expenditures on equipment and structures in manufacturing by \$6.7 billion (1954 dollars) over the 1954–63 period; but for $\sigma=0.2$, which is the value giving the best fit for equipment and structures combined, Table 2 reveals that tax policy increased investment by only \$2.0 billion. By assuming that σ is unity, Hall and Jorgenson have greatly exaggerated the impact of tax policy.

II. *Neoclassical Theory and the Estimation of Tax Policy Effects*

A full understanding of the Hall-Jorgenson model requires careful examination of the neoclassical theory of the firm. Consider a competitive firm that produces output according to the production function $Q=f(L, K)$ and faces prices p, w, and c for Q, L, and K, respectively (c is, of course, an implicit rental price). To maximize profits, the firm should equate the marginal product of each factor to its real price:

$$\frac{\partial Q}{\partial L} = \frac{w}{p} \tag{10}$$

$$\frac{\partial Q}{\partial K} = \frac{c}{p}\,. \tag{11}$$

Equations (10) and (11) form a determinate system (given proper curvature conditions on the function f) which can be solved for the optimum labor and capital inputs, i.e., for the demand functions for labor and capital. Let us assume that the production function is Cobb-Douglas with diminishing returns to scale:

$$\begin{aligned} Q = AL^{\beta}K^{\alpha}, \qquad &\alpha, \beta, A > 0, \\ &0 < \alpha + \beta < 1. \end{aligned} \tag{12}$$

The simultaneous solution of (10) and (11) then yields (13) and (14), L^* and K^* being

of the constant term is a misspecification. Inclusion of a constant term in these relations did not improve their fit greatly.

[9] An important determinant of the uncorrected R^2's in these relations is the correlation between N_t and N_{t-1}. The simple correlation between equipment expenditures and lagged equipment expenditures, for example, is .51 for the period fitted by Hall and Jorgenson; however, this correlation is only .23 for the 1931–41 period and .17 for the 1951–63 period which I fitted.

[10] Jorgenson explicitly assumes diminishing returns in [7, p. 53]. With constant returns, the optimum labor and capital inputs would be either zero or infinite.

$$Q^* = A^{1/(1-\alpha-\beta)}[\beta(p/w)]^{\beta/(1-\alpha-\beta)}[\alpha(p/c)]^{\alpha/(1-\alpha-\beta)}, \tag{15}$$

$$\begin{aligned} K^* = \alpha(p/c)Q^* &= \alpha(p/c)A^{1/(1-\alpha-\beta)}[\beta(p/w)]^{\beta/(1-\alpha-\beta)}[\alpha(p/c)]^{\alpha/(1-\alpha-\beta)} \\ &= A^{1/(1-\alpha-\beta)}[\beta(p/w)]^{\beta/(1-\alpha-\beta)}[\alpha(p/c)]^{(1-\beta)/(1-\alpha-\beta)}, \end{aligned} \tag{17}$$

the optimum inputs of labor and capital.[10] These are the neoclassical demand functions for this problem, and it should be noted that they contain only relative prices as arguments. Substituting L^* and K^* into the production function we obtain (15), that is, desired output is also a function of relative prices.

A test of neoclassical theory in its simplest form would employ (14) as the demand function for capital. As we have already seen, however, Hall and Jorgenson do not use (14); they instead substitute an expression for the marginal product of capital into (11), solve for K, and write

$$K^* = \alpha(p/c)Q. \tag{16}$$

What is missing in this equation is a * on Q; output is a decision variable for the firm, and equation (16) with Q replaced by Q^* is a valid relation between desired capital stock and desired output. It should be written as (17), which is exactly the same as equation (14).

Equation (16), in which Q is actual output, may legitimately be used in estimating an investment function if one is willing to make either of two assumptions. First, it could be assumed that the firm goes through the following iterative decision process. Each period the firm selects the rates of output and labor input that maximize profits, taking capital stock as fixed at the level on hand at the beginning of the period. The short run profit maximizing rate of output is then inserted into (16) to determine the desired stock of capital. Jorgenson [7, p. 53] [8, p. 249] has appealed to this iterative process to justify the treatment of output as a predetermined variable for the capital stock decision. Thus, the entrepreneur does not simply set L, K, and Q at the levels given by (13), (14), and (15); and his groping for the full, profit maximizing solution leads to the inclusion of actual output in the demand function for capital. This process might well be an accurate description of entrepreneurial decision making, but to claim that the argument is a rationalization of the acceleration principle is certainly misleading and can readily cause confusion. With output in the demand function, and with the label "flexible acceleration principle" attached to the demand function, one can easily slip into the mistaken practice of treating output as an exogenous variable rather than a predetermined variable. This is the slip that Hall and Jorgenson have committed in calculating the effects of tax policy. In adopting (16) as the demand function and ignoring the dependence of Q on c, they take the elasticity of K^* with respect to c to be -1, when in fact it is $-(1-\beta)/(1-\alpha-\beta)$, as we see from (14). With decreasing returns to scale, the expression $(1-\beta)/(1-\alpha-\beta)$ is greater than one; hence, it appears that Hall and Jorgenson would understate the response of K^* to changes in c.

To see how tax policy effects should be computed in their model, let us derive the demand function for capital appropriate to the iterative process just described. Suppose the capital stock on hand at the beginning of period t is K_{t-1}. Solving (10) and (12) for the profit maximizing levels of labor input and output in period t, we obtain

$$L_t = [A\beta(p_t/w_t)]^{1/(1-\beta)}K_{t-1}^{\alpha/(1-\beta)} \tag{18}$$

$$Q_t = A^{1/(1-\beta)}[\beta(p_t/w_t)]^{\beta/(1-\beta)}K_{t-1}^{\alpha/(1-\beta)}, \tag{19}$$

and the desired capital stock is

$$K_t = \alpha(p_t/c_t)Q_t, \tag{20a}$$

which may be written as (20b).

$$K_t = \alpha(p_t/c_t)A^{1/(1-\beta)}[\beta(p_t/w_t)]^{\beta/(1-\beta)}K_{t-1}^{\alpha/(1-\beta)}. \tag{20b}$$

Suppose that a tax policy change occurs in period τ, and let $\tilde{c}_t$, $t=\tau, \tau+1, \cdots$, be the rental price of capital that would have prevailed had there been no policy change. Hall and Jorgenson calculate the change in the desired capital stock attributable to the policy change as

$$\begin{aligned} K_t - \tilde{K}_t &= \alpha(p_t/c_t)Q_t - \alpha(p_t/\tilde{c}_t)Q_t \\ &= \alpha p_t Q_t[(1/c_t) - (1/\tilde{c}_t)], \\ &\qquad t = \tau, \tau+1, \cdots, \end{aligned} \tag{21}$$

that is to say, they assume that Q in the demand function is not affected by the change in the rental price c. But this is true only for period τ. From (19) we see that $Q_{\tau+1}$ depends on K_τ; and since K_τ is affected by the change in the rental price, so is $Q_{\tau+1}$. In the absence of the policy change, and assuming for convenience that there are no lags in the investment process, output in period $\tau+1$ would have been

$$\begin{aligned} \tilde{Q}_{\tau+1} &= A^{1/(1-\beta)}[\beta(p_{\tau+1}/w_{\tau+1})]^{\beta/(1-\beta)} \\ &\quad \cdot \tilde{K}_\tau^{\alpha/(1-\beta)} \\ &= A^{1/(1-\beta)}[\beta(p_{\tau+1}/w_{\tau+1})]^{\beta/(1-\beta)} \\ &\quad \cdot [\alpha(p_\tau/\tilde{c}_\tau)Q_\tau]^{\alpha/(1-\beta)}, \end{aligned} \tag{22}$$

and the effect of tax policy on the desired stock of capital in period $\tau+1$ is given by

$$\begin{aligned} &K_{\tau+1} - \tilde{K}_{\tau+1} \\ &= \alpha(p_{\tau+1}/c_{\tau+1})Q_{\tau+1} - \alpha(p_{\tau+1}/\tilde{c}_{\tau+1})\tilde{Q}_{\tau+1} \\ &= \alpha p_{\tau+1}Q_{\tau+1}[(1/c_{\tau+1}) - (1/\tilde{c}_{\tau+1})] \\ &\quad + \alpha p_{\tau+1}(Q_{\tau+1} - \tilde{Q}_{\tau+1})/\tilde{c}_{\tau+1}. \end{aligned} \tag{23}$$

Thus, we see that Hall and Jorgenson would underestimate the effect of tax policy in period $\tau+1$ by the amount $\alpha p_{\tau+1}(Q_{\tau+1}-\tilde{Q}_{\tau+1})/\tilde{c}_{\tau+1}$. A similar argument applies to each succeeding period. It is obviously impossible to carry out these calculations for the Hall-Jorgenson model, since we do not have estimates of all the parameters in (19). We would either have to estimate (19) or use (20b) instead of (20a) as the demand function for capital in estimating the investment function. Equation (20b) could be used recursively to estimate tax policy effects, that is,

$$\begin{aligned} \tilde{K}_t &= \alpha(p_t/\tilde{c}_t)A^{1/(1-\beta)}[\beta(p_t/w_t)]^{\beta/(1-\beta)} \\ &\quad \cdot \tilde{K}_{t-1}^{\alpha/(1-\beta)}, \\ &\qquad t = \tau, \tau+1, \cdots. \end{aligned} \tag{24}$$

A second, more straightforward assumption that would justify using (16) in estimating the investment function, is that actual output is always equal to desired output. If this is the case, it is again clear that Hall and Jorgenson have incorrectly estimated the effects of tax policy, since they have not taken account of the effect of a change in the rental price of capital on desired output, as given by (15). As long as output is not an exogenous variable for the firm, a change in the rental price of capital will bring about a change in output.

This criticism of the Hall-Jorgenson procedure can perhaps be appreciated best by considering the way in which they estimate the impact of the 1962 depreciation guidelines and investment tax credit. They state that, "The effects of the depreciation guidelines of 1962 are significant, but these effects are confined to investment in equipment" [6, p. 413]; and with regard to the investment credit, they state that " . . . the impact is limited to equipment" [6, p. 410]. What these statements mean is that according to their mistaken way of calculating the impact of tax policy, investment in structures was not affected by these policies. The (stock) demand function for structures in the model is

$$K_S = \alpha_S(p/c_S)Q, \tag{25}$$

where α_S is the elasticity of output with respect to the stock of structures. The 1962 tax policy changes affected only the rental price of equipment, c_E. Since c_E does not appear in (25), and since Q is (mistakenly) treated as exogenous, Hall and Jorgenson conclude that the 1962 policy changes did not affect investment in structures. But the neoclassical demand function for structures would include all prices—p, w, c_S, *and* c_E—as arguments, just as the demand for capital in the two factor case is a function of p, w, and c. A fall in c_E *ceteris paribus*, would on the one

hand decrease the demand for structures, since equipment and structures are substitutable inputs according to the Cobb-Douglas specification; on the other hand it would increase the demand for structures since the profit maximizing level of output would be higher, requiring a larger stock of structures. For the Cobb-Douglas case with decreasing returns to scale, the scale effect exceeds the substitution effect; a fall in c_B will increase the demand for structures.[11] Hall and Jorgenson proceeded as if the elasticity of demand for structures with respect to the user cost of equipment is zero, not on the basis of empirical evidence, but because they misunderstand the meaning of their demand function. The output variable in (25) depends on c_B, among other variables, and their error again lies in treating Q as though it were exogenous.

Once it is clearly understood that output is an endogenous variable for the firm, it is easy to see that the price of output that the firm faces is likely to be affected by a change in the rental price of capital. If each firm in the industry increases its output (supply) in response to a fall in the rental price, there would be excess supply of the commodity and its price p would begin to fall. What this means is that the actual percentage increase in the firm's capital stock given a one percent fall in the rental price of capital is likely to be smaller than $(1-\beta)/(1-\alpha-\beta)$, the elasticity given by (14) which is based on the assumption that c can be varied without changing p. In developing the theory of the firm, we assume for convenience that p is not affected by changes in c; but to maintain this assumption in analyzing actual market data, as Hall and Jorgenson have, is certainly unwarranted.

There may be rationalizations for the inclusion of output as an exogenous variable in the demand function for capital; but if output is exogenous, the firm's decision must be formulated as a cost minimization problem, not as a profit maximization problem, and the appropriate factor demand functions would not be those used by Hall and Jorgenson or those given above for the *CES* case.[12] I have argued this point elsewhere and have provided what I consider to be a tenable marriage of neoclassical theory and the acceleration principle which serves as a basis for my own evaluation of tax incentives using alternative specifications of the investment function [3] [4].

III. *Conclusion*

Hall and Jorgenson have assumed a price elasticity of the demand for capital that seems much too high in the light of evidence from data nearly identical to their own, and this causes them to overestimate seriously the effects of tax policy. Furthermore, they have failed to see that their model implies a reaction of output as well as the rental price of capital to a change in tax policy, and they have neglected to take into account the induced output effect in calculating the investment response to tax policy changes. The bias which this introduces into their estimates of tax policy effects cannot be determined without further theoretical and empirical development of the model; it can be said, however, that this error leads them to draw the unwarranted conclusion that the 1962 tax policy changes had no effect on investment in structures. Neither their results nor their procedures stand up to close scrutiny, and their study must therefore be regarded as inconclusive with respect to the effectiveness of tax incentives for investment.

References

1. K. J. Arrow, H. B. Chenery, B. S. Minhas and R. M. Solow, "Capital-Labor Substitution and Economic Efficiency," *Rev. Econ. Stat.*, Aug. 1961, *45*, 225–50.

[11] If the production function is $Q=AL^\alpha K_E^\beta K_S^\gamma$ with $A,\alpha,\beta,\gamma>0$ and $0<\alpha+\beta+\gamma<1$, the elasticity of demand for structures with respect to the rental price of equipment is $-\beta/(1-\alpha-\beta-\gamma)$. Since $\beta/(1-\alpha-\beta-\gamma)$ is positive, a fall in the rental price of equipment must increase the demand for structures.

[12] Assuming cost minimization, the firm's demand for capital in the case of the Cobb-Douglas function (12) is

$$K^* = A^{-1/(\alpha+\beta)}(\beta/\alpha)^{-\beta/(\alpha+\beta)}(c/w)^{-\beta/(\alpha+\beta)}Q^{1/(\alpha+\beta)},$$

where w is the wage rate. For the *CES* function (2) but with returns to scale equal to θ, the demand function is

$$K^* = [\beta + (1-\beta)^\sigma\beta^{1-\sigma}(c/w)^{\sigma-1}]^{\sigma/(1-\sigma)}Q^{1/\theta}.$$

2. M. Brown and J. S. de Cani, "Technological Change and the Distribution of Income," *Internat. Econ. Rev.*, Sept. 1963, *4*, 289–309.
3. R. M. Coen, "Effects of Tax Policy on Investment in Manufacturing," *Am. Econ. Rev.*, May 1968, *58*, 200–11.
4. ———, *Tax Policy and Investment in Manufacturing, 1954–66*, Ph.D. dissertation, Northwestern Univ. 1967; also to be published by the Brookings Institution in the proceedings of a conference on tax policy and investment held on November 3, 1967.
5. R. Eisner and M. I. Nadiri, "Investment Behavior and Neoclassical Theory," *Rev. Econ. Stat.*, Aug. 1968, *50*, 369–82.
6. R. E. Hall and D. W. Jorgenson, "Tax Policy and Investment Behavior," *Am. Econ. Rev.*, June 1967, *57*, 391–414.
7. D. W. Jorgenson, "Anticipations and Investment Behavior," pp. 35–92 in *The Brookings Quarterly Econometric Model of the United States*, James S. Duesenberry, et al., ed., Chicago 1965.
8. ———, "Capital Theory and Investment Behavior," *Am. Econ. Rev.*, May 1963, *53*, 247–59.
9. ———, "The Theory of Investment Behavior," 129–55 in *Determinants of Investment Behavior*, Robert Ferber, ed., Universities-National Bureau Conference Series, No. 18, New York 1967.
10. ———; and J. A. Stephenson, "The Time Structure of Investment Behavior in United States Manufacturing, 1947–1960," *Rev. Econ. Stat.*, Feb. 1967, *65*, pp. 16–27.
11. L. R. Klein and P. Taubman, "Impact of Accelerated Depreciation and Investment Tax Credit on Investment and General Economic Activity," to be published by The Brookings Institution in the proceedings of a conference on tax policy and investment held at the Institution on November 3, 1967.
12. M. Nerlove, "Recent Empirical Studies of the CES and Related Production Functions," 55–122 in *The Theory and Empirical Analysis of Production*, Murray Brown, ed., National Bureau of Economic Research, Studies in Income and Wealth, No. 31, New York, 1967.
13. K. Sato, "A Two-Level Constant-Elasticity-of-Substitution Production Function," *Rev. Econ. Stud.*, Apr. 1967, *34*, 201–18.
14. H. Uzawa, "Production Functions with Constant Elasticities of Substitution," *Rev. Econ. Stud.*, Oct. 1962, *29*, 291–99.

TAX POLICY AND INVESTMENT BEHAVIOR: COMMENT

By Robert Eisner*

Robert E. Hall and Dale W. Jorgenson, after declaring, "The customary justification for the belief in the efficacy of tax stimulus does not rely on empirical evidence" [3, p. 391], have presumed, upon examination, to see in the evidence "little room for doubt about the efficacy of tax policy in influencing investment behavior" [p. 413]. It is the purpose of this comment to demonstrate: (1) the conclusions by Hall and Jorgenson are based on assumption, not empirical evidence; (2) analysis of the empirical data, including basic data made available by Jorgenson, do little to confirm the " . . . article of faith among both policymakers and economists" regarding " . . . effectiveness of tax policy in altering investment behavior" [p. 391].

I. *The Work by Hall and Jorgenson*

Hall and Jorgenson claim to "have calculated the effects of changes in tax policy and investment behavior" for the accelerated depreciation in the Internal Revenue Code of 1954, the reduction in tax depreciation brought on by the new "guidelines" in 1962, and the investment tax credit for equipment and machinery, also in 1962. In addition, they consider the hypothetical effects of the ultimate liberalization of tax depreciation: the first year write-off or expensing of invest-

* The author is professor of economics, Northwestern University. He has benefitted considerably from the comments of Bert G. Hickman as well as from the financial support of the National Science Foundation and the Ford Foundation for research in investment.

ment expenditures. Their substantive conclusions—and they are substantive indeed—include the assertions that from 6.8 to 11.4 percent of various categories of gross investment in the 1954–63 period resulted from the accelerated depreciation introduced in 1954 (p. 408), that 14.8 to 17.6 percent of the *net* investment in equipment in 1963 was due to the change in guidelines, and that 40.9 to 48.6 percent of net investment or just over 10 percent of gross investment in 1963 could be attributed to the investment tax credit [p. 410]. I shall demonstrate that these conclusions stem directly from the assumptions of the model used by Hall and Jorgenson and have little or nothing to do with the data.

It is important to make clear just wherein the work by Hall and Jorgenson, which does involve estimates of investment functions from annual data of the U.S. economy for the years 1931–41 and 1950–63, fails to offer empirical evidence on the role of tax policy. To do so, we must consider first the assumptions or constraints of Jorgenson's "neoclassical theory of optimal capital accumulation" [pp. 391–92].

As should be well known to readers of various recent papers by Jorgenson and others, the assumption of a Cobb-Douglas production function implies that the desired level of capital K^* is:

$$K^* = \alpha \frac{pQ}{c}, \tag{1}$$

"where p is the price of output, Q is quantity, c the rental price of capital, and α the *constant* elasticity of output with respect to capital" [p. 396]. Then, assuming that net investment is a distributed lag function of previous changes in the desired stock of capital and that replacement investment is proportional to existing capital stock, we derive

$$I_t = \sum_{s=0}^{\infty} \mu_s' \Delta K_{t-s}^* + \delta K_t, \tag{2}$$

where I_t is gross investment.

With the replacement coefficient, δ, estimated extraneously, Hall and Jorgenson utilize a Koyck, geometric series form of their general Pascal lag estimator for their regression function in net investment:

$$N_t = \alpha\gamma_0' \Delta \frac{p_t Q_t}{c_t} + \alpha\gamma_1' \Delta \frac{p_{t-1}Q_{t-1}}{c_{t-1}} + \omega' N_{t-1} + \epsilon_t. \tag{3}$$

They then proceed to estimate the parameters of (3), taking

$$c = q(r + \delta) \frac{(1 - k)(1 - uz)}{1 - u}, \tag{4}$$

where q is the price of capital goods, r is the discount rate, δ the rate of replacement, k the rate of the investment tax credit, u the business income tax rate, and z the present value of the depreciation deduction on one dollar's investment.[1] The depreciation method, which affects the present value of the tax depreciation deduction, and the investment tax credit thus enter into the measure of the rental price of capital, c. What is critical, however, is that Hall and Jorgenson (as in the papers by Jorgenson, and Jorgenson and Stephenson, [4], [6]) do not offer an independent or unconstrained estimate of the parameter of their measure of the "rental price of capital," let alone independent or unconstrained estimates of the role of the particular components of the price of capital affected by tax policy. Hall and Jorgenson rather have constrained their estimator in (3) so that given percentage increases in p and Q, and decreases in c, must all have identical effects on investment regardless of anything in the empirical data.

Thus, if variations in c, elsewhere referred to by Jorgenson as "the user cost of capital,"[2] and a fortiori variations in the tax component of that cost, really had no effect whatsoever on investment, but p or more

[1] The current Hall and Jorgenson usage of z is identical to the correction offered several years ago by Robert M. Coen, in [1]. Elsewhere, however, papers by Jorgenson, and Jorgenson and Stephenson, have used the ratio of depreciation charges to replacement at current prices instead of z as here defined. (See [4, p. 248] [6, p. 17, fn. 7], for example.)

[2] See [4, p. 249], for example.

likely Q, output, did have some substantial effect on investment, we could still expect positive regression coefficients of $\Delta(pQ/c)$. These might reflect entirely the effects of changes in output, ΔQ (the role of the accelerator), and be biased downward by the errors in the variables introduced by variance in c.[3] The positive regression coefficients so derived would offer no evidence of any influence on investment stemming from the user cost of capital as opposed to the influence of output.

There is indeed internal evidence in the article by Hall and Jorgenson (as well as in other work by Jorgenson and associates) that precisely this is happening. If the production function is Cobb-Douglas, and the other maximization and implicit expectational and informational assumptions of the model are met, then it is true that, on substituting (1) in (2), the distributed lag coefficients, μ'_s, would sum to unity, although this is an implication of the model and its assumptions and not a "fact" as Hall and Jorgenson assert [p. 400]. The implications of this, since the regression coefficients of $\Delta(pQ/c)$ are so low, are that the estimated elasticities of output with respect to capital, $\hat{\alpha}$, are low, particularly in manufacturing. There the sum of the elasticities for equipment and structures is approximately .11,[4] well below the .25 to .33 usually found in direct estimates of this function (or the .35 elsewhere hailed by Jorgenson as confirmed by the data).[5]

It is to be noted that the assumption that the sum of the $\mu'_s = 1$, coupled with the assumption that the coefficients of appropriate measures of changes in the p/c and Q components of $\Delta(pQ/c)$ would be identically equal to $\alpha\gamma'_0$ and $\alpha\gamma'_1$ if these coefficients were estimated separately, are the basis for all of the conclusions by Hall and Jorgenson about the effects of various tax policies. For given these estimates of $\alpha\gamma'_0$ and $\alpha\gamma'_1$, and ω', and hence of $\mu'_s\alpha$, where α as always in the Jorgenson model is a parameter, it becomes a matter of cumbersome arithmetic —or simple computer calculations—to work out the effects on investment of changes in depreciation or tax policy in terms of their calculated effects upon c and hence upon p/c or pQ/c, at least if one goes along with the further implicit assumption that p and Q are unaffected by changes in c.

The estimates by Hall and Jorgenson of the effects of taxes on investment are all based upon their estimates of the coefficients of Δc_t and Δc_{t-1} although these coefficients are never specifically estimated. Rather they are taken bodily from the coefficients of $\hat{\alpha}\Delta(p_tQ_t/c_t)$ and $\hat{\alpha}\Delta(p_{t-1}Q_{t-1}/c_{t-1})$. This is done despite the evidence of downward bias in the estimates of α, cited above, which suggests errors in variables or misspecification of the relation. What is more, there is no empirical estimate that all of the various components of c specified in (4) contribute to investment in the way that Hall and Jorgenson have concluded. Hall and Jorgenson assume that the elasticity of desired capital, K^*, with respect to rental price of capital, c, is unity, and use regression coefficients of a variable that includes price of capital goods and the rate of output to estimate only the lag distribution of desired to actual capital stock. Hall and Jorgenson then further assume the role of the tax parameters, k, u, and z, and the cost of capital, r, is that specified by (4). It is not my purpose here to discuss the plausibility of the theory; it is imperative to distinguish theory from empirical findings. Hall and Jorgenson offer none of the latter which bears on the role of taxes in investment.

II. *Another Look At Empirical Data*

While Hall and Jorgenson have not used their empirical data to measure the role of the rental price or user cost of capital, c, on which their conclusions about taxes and investment depend, quarterly data made available by Jorgenson[6] have been used by Eisner and Nadiri [2] to estimate separately the roles of relative prices, p/c, and output, Q. This analysis, applies to a log-linear relation the general Pascal lag estimator, of which the

[3] See [2, p. 375, fn. 26].
[4] [3, Table 2, p. 400].
[5] See [5, esp. p. 471].

[6] These are described briefly in the Appendix.

estimator employed by Hall and Jorgenson is a special case. Eisner and Nadiri offer evidence that variations in output, not in the user cost of capital, are most associated with investment. Indeed, in quarterly data using the best fitting Koyck regression, involving seven quarterly lag weights γ_i, the elasticity of capital stock with respect to relative prices (p/c) was estimated at only .1576 while that with respect to output was .8158 [2, Table 2].[7] Eisner and Nadiri confirmed that a downward bias is introduced into estimates of the γ_i by misspecification of the role of c; they found uniformly lower estimates of elasticities of capital stock with respect to the combined variable, pQ/c; the highest estimate there was .464 and for the best fitting regression it was .245.[8]

The calculations by Hall and Jorgenson of the effects of changes in the user cost of capital brought about by changes in tax policies are based upon the assumption that the elasticity of capital stock with respect to the user cost of capital is unity. It must be realized that the unconstrained estimates of this elasticity in the neighborhood of .16 provided by Eisner and Nadiri, using Jorgenson's data, suggest that if the other assumptions and arguments built into the paper by Hall and Jorgenson are correct, their estimates of the effect of tax policy should be divided roughly by six.

A further direct test of the role of tax policy has been carried out by utilizing two sets of user cost variables made available by Robert M. Coen from quarterly data from 1947 to 1966, calculated on essentially the basis indicated by Hall and Jorgenson. One set of user cost variables is prepared on the assumptions of straight line depreciation with no change in guideline life or tax credit over the period. The other set of user cost variables encompasses the various changes in 1954 and 1962 considered by Hall and Jorgenson, namely the introduction of sum-of-years-digits depreciation in 1954, and of guideline depreciation and the investment credit on equipment in 1962. While the annual data for manufacturing and nonmanufacturing utilized by Hall and Jorgenson have not been available to me, I have added Coen's measures of the effects of tax policy on user cost to the quarterly manufacturing data used by Jorgenson and Stephenson, reported upon by Eisner and Nadiri [2]. The following equations were then estimated:

$$\Delta\, ln\, K_t = \sum_{i=1}^{7}\left[\gamma_i \Delta\, ln \left(\frac{pQ}{c}\right)_{t-i} + \gamma_{\Delta ci}(\Delta\, ln\, c^0 - \Delta\, ln\, c^x)_{t-i}\right] + \sum_{j=1}^{r} \omega_j \Delta\, ln\, K_{t-j} + v_5 \qquad r = 1, 2, 3 \tag{5}$$

$$\Delta\, ln\, K_t = \sum_{i=m}^{m+4}\left[\gamma_{pi}\Delta\, ln\left(\frac{p}{c}\right)_{t-i} + \gamma_{\Delta ci}(\Delta\, ln\, c^0 - \Delta\, ln\, c^x)_{t-i} + \gamma_{Qi}\Delta\, ln\, Q_{t-i}\right] + \sum_{j=1}^{r}\omega_j \Delta\, ln\, K_{t-j} + v_6 \qquad m = 1, 2, 3;\ r = 1, 2 \tag{6}$$

where

c^0 = user cost with straight-line depreciation, useful life of 18 years, no allowance for tax credit or tax cuts,

and

c^x = user cost with various tax incentives: sum-of-years-digits tax depreciation beginning in 1954, useful life reduced to 13 years and tax credit and tax cuts included from 1962–III.

[7] The estimating equations corresponded respectively to (5) and (6) without the terms involving c^0 and c^x. Similar sets of estimates were obtained from arithmetic forms of the relation, as indicated in [2, pp. 371–72, fn. 19].

[8] Errors in variables do not, of course, always bias estimates of parameters downward. As explained in [2, p. 375, fn 26], however, the particular relation among covariances and variances of errors and variables which might prevent such a downward bias is not to be found in the body of data with which we are concerned.

The unconstrained estimate of elasticity of capital with respect to any of the independent variables, h, in (5) or (6), may then be written

$$\hat{E}_h = \frac{\sum_i \gamma_{hi}}{1 - \sum_j \omega_j}. \tag{7}$$

The rationale of this formulation may be seen by noting that we are including first Jorgenson's user cost variable,[9] c, described below in (8). Over the years of our sample this Jorgenson c reflects the effect of the tax incentives only in the role of v, the ratio of tax depreciation charges to current replacement cost, in the expression

$$c = q\left[\left(\frac{1 - uv}{1 - u}\right)\delta + \left(\frac{1 - uw}{1 - u}\right)r\right], \tag{8}$$

where w, which need not here detain us, is the proportion of the cost of capital, r, deductible for tax purposes, and the other symbols are as defined previously. With v involving the slowly moving total of depreciation charges stemming from capital acquisitions of many past years, the measure of c defined in (8), compared to the measure of c in (4), used by Jorgenson and Hall and by Coen, captures relatively little of the effect of accelerated depreciation in the years 1954–62. Hence, we may take Jorgenson's c as an approximation of c^0, use $\Delta\, ln\, c^0 - \Delta\, ln\, c^x$ in (5) and (6) in order to reflect the role of the change in user cost, or differences in changes in user cost brought on by tax incentives, and write the equilibrium relation underlying (6) as

$$\begin{aligned} K^* &= \beta\left(\frac{p}{c}\right)^{E_p}\left(\frac{c}{c^x}\right)^{E_{\Delta c}} Q^{E_q} \\ &= \beta \frac{p^{E_p}}{c^{(E_p - E_{\Delta c})} c^{x(E_{\Delta c})}} Q^{E_q}. \end{aligned} \tag{9}$$

If the role of tax incentives were correctly specified in c^x we should have $E_{\Delta c} = E_p$ and hence

$$K^* = \beta\left(\frac{p}{c^x}\right)^{E_p} Q^{E_q} \tag{10}$$

which, in Jorgenson's special Cobb-Douglas case with $E_{\Delta c} = E_p = E_q = 1$, would reduce to (1). Thus a value of $E_{\Delta c} = 1$ would confirm the specified role at least of accelerated depreciation, which was the major investment tax incentive in our sample period. Estimates of $E_{\Delta c}$ greater than zero, but less than unity, would suggest some positive role for tax incentives in stimulating investment, but not that specified by Hall and Jorgenson. Estimates of $E_{\Delta c} \leqq 0$ would indicate that this formulation has found no discernible effect of the tax incentives (in this instance essentially accelerated depreciation in the 1954–62 period) in stimulating investment.

Table 1, presenting estimates from the best fitting regressions of equation (5), where the rental price of capital, c_1, is based upon a "cost of capital," r_1, defined as interest and earnings divided by the value of all outstanding securities, offers an immediate empirical contradiction of the conclusions by Hall and Jorgenson. The sum of the coefficients of the tax incentive variable, while not differing significantly from zero, is actually negative, equaling $-.0308$. Further, the sum of the coefficients of cost-deflated output, $\Delta\, ln\, (pQ/c_1)$, while significantly positive at a .05 probability level, is so low that the unconstrained estimate of the elasticity of capital stock to what is referred to in the Jorgenson models as "the desired level of capital" is only .2778. This is further evidence of an underlying misspecification in (1), introducing errors in measurement of desired capital which bias the estimated coefficients toward zero.

The results presented in Table 2, where the bond yield, r_2, is used as a measure of the cost of capital, are essentially the same. Again the sum of the coefficients of the tax factor is negative, this time equaling $-.0235$ but not significantly different from zero. Again the estimated elasticity of capital stock with respect to cost deflated output is inappropriately low, equaling .2970.

Estimates of equation (6), presented in Table 3, offer a substantiation of the Eisner-Nadiri findings that it is changes in output rather than in relative prices which seem to account most for investment expenditures.

[9] See [7, p. 186].

TABLE 1—CAPITAL EXPENDITURES, CHANGE IN COST-DEFLATED OUTPUT AND TAX FACTOR, TOTAL MANUFACTURING QUARTERLY DATA, 1947–62, USER COST OF CAPITAL (c_1) INVOLVING INTEREST AND EARNINGS[a]

$$\Delta \ln K_t = \alpha' + \sum_{i=1}^{7} \left[\gamma_i \Delta \ln \left(\frac{pQ}{c_1}\right)_{t-i} - \gamma_{\Delta ci}(\Delta \ln c^0 - \Delta \ln c^*)_{t-i} \right] + \sum_{j=1}^{2} \omega_j \Delta \ln K_{t-j} + v$$

(1) i, j	(2) γ_i	(3) $\gamma_{\Delta ci}$	(4) ω_j
1	.0042 (.0036)	.0179 (.0143)	1.4671 (.1336)
2	.0015 (.0035)	−.0034 (.0235)	−.5408 (.1266)
3	.0067 (.0041)	−.0233 (.0233)	
4	.0083 (.0039)	−.0275 (.0232)	
5	.0096 (.0041)	−.0166 (.0235)	
6	−.0025 (.0045)	.0352 (.0236)	
7	−.0073 (.0043)	−.0131 (.0221)	
Sums	.0205 (.0100)	−.0308 (.0613)	.9263 (.0423)
$\hat{E}$[b]	.2778	−.4172	

[a] $\bar{R}^2 = .933$; $N = 56$; $\alpha' = .0002$. (.0003)

[b] $\hat{E}$ = estimated elasticity of capital with respect to variables associated with indicated parameters.

TABLE 2—CAPITAL EXPENDITURES, CHANGE IN COST-DEFLATED OUTPUT AND TAX FACTOR, TOTAL MANUFACTURING QUARTERLY DATA, 1947–62, USING BOND RATE IN USER COST OF CAPITAL (c_2)[a]

$$\Delta \ln K_t = \alpha' + \sum_{i=1}^{7} \left[\gamma_i \Delta \ln \left(\frac{pQ}{c_2}\right)_{t-i} + \gamma_{\Delta ci}(\Delta \ln c^0 - \Delta \ln c^*)_{t-i} \right] + \sum_{j=1}^{2} \omega_j \Delta \ln K_{t-j} + v$$

(1) i, j	(2) γ_i	(3) $\gamma_{\Delta ci}$	(4) ω_j
1	.0098 (.0045)	.0010 (.0165)	1.5908 (.1265)
2	.0062 (.0046)	−.0069 (.0253)	−.6754 (.1252)
3	.0003 (.0042)	−.0027 (.0231)	
4	−.0004 (.0040)	−.0202 (.0238)	
5	.0053 (.0038)	−.0337 (.0247)	
6	.0035 (.0046)	.0334 (.0248)	
7	.0004 (.0050)	.0056 (.0226)	
Sums	.0251 (.0132)	−.0235 (.0701)	.9154 (.0408)
$\hat{E}$	.2970	−.2774	

[a] $\bar{R}^2 = .924$; $N = 56$; $\alpha' = .0004$. (.0003)

Further they contradict the hypothesis that changes in the tax structure have had a discernible effect in stimulating investment. It will be noted in Table 3 that the sum of the $\gamma_{pi} = .0195$, with a standard error of .0092 indicating statistical significance at at least a .05 level. This implies an elasticity of capital stock with respect to relative prices, including a cost of capital involving interest and earnings, of .2274. The sum of the γ_{qi} of .0632, with a standard error of .0127, indicates both a clearly statistically significant role for changes in output and an estimated elasticity of capital stock with respect to output of .7385, not too far out of line with the notion that capital stock eventually moves roughly in proportion to output. The substantial difference between these two elasticities, .2274 for relative prices and .7385 for output, is further evidence of the misspecification of unity as the common elasticity of capital stock with respect to prices, user cost, and output, to be found in

TABLE 3—CAPITAL EXPENDITURES, CHANGE IN OUTPUT, RELATIVE PRICES AND TAX FACTOR, TOTAL MANUFACTURING QUARTERLY DATA, 1947–62, USER COST OF CAPITAL (c_1) INVOLVING INTEREST AND EARNINGS[a]

$$\Delta \ln K_t = \alpha' + \sum_{i=1}^{5} \left[\gamma_{pi} \Delta \ln \left(\frac{p}{c_1}\right)_{t-i} + \gamma_{\Delta ci}(\Delta \ln c^0 - \Delta \ln c^x)_{t-i} + \gamma_{qi} \Delta \ln Q_{t-i} \right] + \sum_{j=1}^{2} \omega_j \Delta \ln K_{t-j} + v$$

(1) i, j	(2) γ_{pi}	(3) $\gamma_{\Delta ci}$	(4) γ_{qi}	(5) ω_j
1	.0030 (.0035)	.0159 (.0137)	.0188 (.0065)	1.2008 (.1167)
2	.0018 (.0034)	.0024 (.0245)	.0069 (.0067)	−.2864 (.1192)
3	.0019 (.0041)	.0023 (.0247)	.0085 (.0060)	
4	.0038 (.0041)	−.0119 (.0245)	.0081 (.0058)	
5	.0090 (.0043)	−.0226 (.0219)	.0210 (.0056)	
Sums	.0195 (.0092)	−.0139 (.0516)	.0632 (.0127)	.9144 (.0384)
$\hat{E}$	.2274	−.1622	.7385	

[a] $\hat{R}^2 = .934$; $N = 58$; $\alpha' = -.0002$. (.0004)

all of the work on investment with which Jorgenson has been associated.[10] What is more, however, for our own immediate purpose, we may note that the sum of the $\gamma_{\Delta ci}$ is again not significantly different from zero, and again negative, amounting to $-.0139$.

Table 4 is similar in form to Table 3 except for the use of the bond rate as the cost of capital instead of the ratio of interest and earnings to the value of securities. Finally, those looking for empirical support for a role of the tax factor in investment may find some small comfort. For while again it is only changes in output which prove statistically significant, and relative prices actually show up on balance with a negative role, the sum of the $\gamma_{\Delta ci}$ is positive and implies an elasticity of capital stock of .62 with respect to changes in user cost brought about by changes in depreciation and tax policy. I trust, however, that readers will find the empirical support here for the role of tax incentives to be scant, in that the

[10] As pointed out with regard to similar results in [2], these findings are consistent with the implications of a first degree *CES* production function, where

$$K^* = \beta \left(\frac{p}{c}\right)^{E_p} Q,$$

but not with its special case, the Cobb-Douglas function, since E_p, the elasticity of substitution, is clearly less than unity. If one could assume $\partial Q/\partial K = c/p$ even with nonconstant returns, thus forcing a contradiction in other marginal conditions, one could generalize the above for a *CES* function of degree v, to

$$K^* = \beta \left(\frac{p}{c}\right)^{\sigma} Q^{\sigma + (1-\sigma)/v},$$

where, for $E_p = \sigma < 1$ and $v > 1$,

$$E_q = \sigma + \frac{1-\sigma}{v} = 1 + \frac{(1-\sigma)(1-v)}{v} < 1,$$

which is consistent with our estimates.

TABLE 4—CAPITAL EXPENDITURES, CHANGE IN OUTPUT, RELATIVE PRICES AND TAX FACTOR, TOTAL MANUFACTURING QUARTERLY DATA, 1947–62, USING BOND RATE IN USER COST OF CAPITAL (c_2)[a]

$$\Delta\, ln\, K_t = \alpha' + \sum_{i=1}^{5}\left[\gamma_{pi}\Delta\, ln\left(\frac{p}{c_2}\right)_{t-i} + \gamma_{\Delta ci}(\Delta\, ln\, c^0 - \Delta\, ln\, c^s)_{t-i} + \gamma_{qi}\Delta\, ln\, Q_{t-i}\right] + \sum_{j=1}^{2}\omega_j\Delta\, ln\, K_{t-j} + v$$

(1) i, j	(2) γ_{pi}	(3) $\gamma_{\Delta ci}$	(4) γ_{qi}	(5) ω_j
1	.0001 (.0052)	.0189 (.0166)	.0219 (.0056)	1.2567 (.1250)
2	−.0015 (.0054)	.0243 (.0305)	.0089 (.0062)	−.3487) (.1240)
3	−.0060 (.0064)	.0340 (.0307)	.0093 (.0059)	
4	−.0088 (.0066)	.0068 (.0300)	.0025 (.0056)	
5	.0016 (.0068)	−.0270 (.0231)	.0148 (.0053)	
Sums	−.0146 (.0141)	.0570 (.0647)	.0574 (.0130)	.9080 (.0387)
$\hat{E}$	−.1585	.6200	.6235	

[a] $\hat{R}^2 = .924$; $N = 58$; $\alpha' = -.0002.$ $(.0004)$

sum of these $\gamma_{\Delta ci}$ does not differ from zero by even as much as its standard error, and the positively estimated elasticity of capital stock with respect to the tax factor is not statistically significant at any reasonable probability level.[11]

Experimentation with other lag structures, within the limits of the capacity of our regression program, in no case revealed any greater evidence of the role of the tax factor than that exhibited in Table 4. Lags of the independent variables running from 2 to 6, 3 to 7, 4 to 8, and 5 to 9, with two lagged values of the dependent variable, indicated partial elasticities of capital stock with regard to the tax variable which were negative in five out of ten cases, including those of the three best fitting regressions. In no case was the sum of regression coefficients of the tax factor greater than its standard error, or as great as the figure of .0570 shown in Table 4. And in no case was the estimated elasticity as great as that indicated, without statistically significant difference from zero, in Table 4.

[11] It should be recognized, however, that failure to find a statistically significant positive elasticity with respect to any variable is not the same as rejecting the hypothesis that the elasticity is positive. Even where the estimated elasticity with respect to the tax incentive factor was negative, standard errors were too high to reject the hypothesis of a positive elasticity at as much as the .05 probability level. Similarly, only in Table 4 do results indicate clear rejection of the hypothesis of unitary positive elasticity with respect to the relative price variable, although the more precise estimates of elasticities in Eisner and Nadiri [2] do quite generally point to price elasticities significantly below unity.

III. *Conclusion*

Hall and Jorgenson, contrary to the implications and conclusions of their article, have offered no empirical evidence of the role of

tax factors in investment. Their specification of the investment function with the constraint that parameters of the price of output, output, and the reciprocal of the rental price or user cost of capital must all be identical, along with their specification of an exact relation between the user cost of capital and the various tax incentives they consider, make it impossible for them to disentangle any possible influence of these tax incentives on investment. There is hence no justification for the quantitative estimates of these influences which they present in their article.

There is internal evidence in their findings from annual data of the inter-war and post-World War II periods that their investment function is in fact misspecified. When a revised form of their basic investment function, permitting separate estimation of parameters of relative prices and output, is confronted with Jorgenson's quarterly data for manufacturing, this misspecification is confirmed. The role of relative prices, on which the influence of tax incentives depends critically in the Hall-Jorgenson article, is found to be quite small, at most about 1/6 that assumed by Hall and Jorgenson. More specifically, when a variable measuring the changes in the user cost of capital brought about by tax measures, chiefly accelerated depreciation, is introduced specifically into the relation, the best estimates of the effect of the variable are negative and no statistically significant role for the tax incentive is found in any of the estimates.

Appendix

The following quarterly data, generally from 1947-I through 1962-IV were obtained from Jorgenson:

1. Gross investment from the OBE-SEC Quarterly Investment Surveys (reported in the *Survey of Current Business*), price deflated.

2. Capital stock net of depreciation at the beginning of 1948 and 1960 from the OBE Capital Goods Study (G. Jaszi *et al.* in *Survey of Current Business*, November 1962) with intervening quarterly values along with the rate of depreciation δ, calculated from the recursive relation,

$$K_{t+1} = I_t + (1 - \delta)K_t.$$

3. Gross value added—pQ—the sum of corporate and noncorporate profits before taxes, depreciation, net monetary interest, and total compensation of employees. These data for total manufacturing were allocated among quarters in proportion to the corresponding series in the FTC-SEC *Quarterly Financial Reports*.

4. A first measure of the rental price or user cost of capital, c_1, based on r_1, defined as (corporate profits after taxes plus net monetary interest)/the value of all outstanding securities. The value of equity was estimated as the ratio of corporate profits after taxes to the earnings-price ratio for manufacturing corporations reported by Standard and Poor's. The value of debt is estimated as the ratio of net monetary interest to the bond yield for manufacturing corporations.

5. A second measure of the rental price or user cost of capital, c_2, based on r_2, the U.S. Government Long-Term Bond Rate as reported in the *Survey of Current Business*. All of these data are reported to be seasonally adjusted and measured at quarterly rates. All price indexes take the value of unity in 1954.

Coen's measures of user cost with and without tax incentives are described in further detail in [1]. The only additional variable, introduced by Eisner and Nadiri, is the wholesale price index, also given the value of unity in 1954. This index was considered to be p and output of each quarter, Q_t, was then calculated by dividing the Jorgenson figure, p_tQ_t, by our p_t. Our measure of relative prices was p/c.

Fuller accounts of Jorgenson's data are to be found in various of the references cited but particularly in the "Statistical Appendix" [7, pp. 76–77] of the manuscript copy.

References

1. R. M. Coen, "Tax Policy and Investment in Manufacturing, 1954–1966," in The Brookings Institution, *Studies of Government Finance*, Washington 1967.
2. R. Eisner and M. I. Nadiri, "On Investment Behavior and Neo-Classical

Theory," *Rev. Econ. Stat.*, Aug. 1968, *50*, 369–82.
3. R. E. HALL AND D. W. JORGENSON, "Tax Policy and Investment Behavior," *Am. Econ. Rev.*, June 1967, *57*, 391–414.
4. D. W. JORGENSON, "Capital Theory and Investment Behavior," *Am. Econ. Rev.*, May 1963, *53*, 247–59.
5. ———, "Comment" on Eisner, "Capital and Labor in Production: Some Direct Estimates," in M. Brown, ed., *The Theory and Empirical Analysis of Production*, NBER Studies in Income and Wealth, Vol. 31. New York 1967.
6. ——— AND J. A. STEPHENSON, "The Time Structure of Investment Behavior in United States Manufacturing, 1947–1960," *Rev. Econ. Stat.*, Feb. 1967, *49*, 16–27.
7. ——— AND J. A. STEPHENSON, "Investment Behavior in U.S. Manufacturing, 1947–60," *Econometrica*, Apr. 1967, *35*, 169–220.

TAX POLICY AND INVESTMENT BEHAVIOR: REPLY AND FURTHER RESULTS

By ROBERT E. HALL AND DALE W. JORGENSON*

In our earlier paper in this *Review*,[1] we presented the results of an empirical study of the impact of changes in American tax policy on investment expenditures. The basic principles of the investment function underlying our work are the familiar ones: starting from a hypothesis about the form of the production function, we derived the profit-maximizing demand for capital input as a function of output and the relative price of capital services. We then estimated the parameters of the lagged adjustment of actual capital to this desired level. One property of the demand function, namely its elasticity with respect to the price of capital services, obviously has a crucial role in the application of such an investment function to the measurement of the effects of tax policy. We chose the relatively simple Cobb-Douglas parametrization for the production function. This choice implied the particular numerical value of unity for the own-price elasticity of demand for capital (or elasticity of substitution). At the time we made the choice, we believed that it was a reasonable one in the light of evidence from studies of production functions and factor demand. Evidence accumulated since then has strongly confirmed our hypothesis, as we will show here.

Robert Coen [4] has correctly pointed out that in our original exposition we did not provide a detailed justification for the choice of this hypothesis. Our first objective is to review the extensive evidence on the elasticity of substitution now available. The elasticity of substitution is difficult to estimate from time series since it is essentially a second order parameter in the relationship between products and factors, as Kmenta [18] and Nelson [21] have pointed out. As we observe in Section I, relatively minor errors in specification or seemingly unimportant differences in methods of measurement may have a substantial impact on the estimated elasticity of substitution. The elasticity can be estimated more reliably from individual cross sections and from successive cross sections, as Griliches [9] has demonstrated. New evidence made available by Zarembka [26] reinforces Griliches' conclusions and removes important sources of ambiguity in the interpretation of earlier studies.

Our second purpose is to present a new set of estimates of the time form of the response of investment to changes in its determinants. These estimates, based on a refinement of our previous econometric method, suggest that the impact of tax policy is considerably more rapid than we originally estimated. Our earlier estimates of the timing of changes in investment following a change in tax policy need revision in light of this finding. Our new results are in much better agreement with the results of studies of the time

* The authors are members of the faculty of the University of California, Berkeley.

[1] See [10] in this *Review*, June 1967.

structure of investment behavior at the level of industry groups or for individual firms.

Our third objective is to extend our previous estimates to measure the impact of policy changes since 1963 on investment expenditures. Since 1963 several important changes in tax policy have been carried out; these changes provide a further opportunity for evaluation of the responsiveness of investment expenditures to tax policy. We first consider the effects of reduction of the rate of taxation on corporate profits from 52 percent to 48 percent in 1964. This reduction was combined with elimination of the requirement that the base for depreciation charges be reduced by the amount of the investment tax credit, making the tax credit adopted in 1962 considerably more effective. We find that the main stimulus to investment resulting from the 1964 tax cut came from changes in the investment tax credit rather than changes in the tax rate. If the tax rate alone had been reduced with no alteration in the investment tax credit, the direct impact of the tax cut of 1964 on the level of investment would have been small and negative. A second change in tax policy we analyze is the suspension of the investment tax credit and certain types of accelerated depreciation in late 1966 and early 1967. These changes had a substantial impact on investment and would have had a much greater impact if the suspension had been kept in force for the period stipulated at the time of the suspension.

I. *The Elasticity of Substitution*

In our econometric model of investment behavior we have maintained the hypothesis that the elasticity of substitution is equal to unity. Although relatively minor errors in specification or small differences in methods of measurement have substantial impact on estimates of the elasticity of substitution from time series, this parameter can be estimated reliably from cross sections, either through the study of factor demand or from direct estimates of the production function. We first review estimates of the elasticity of substitution from empirical studies of factor demand. Our maintained hypothesis is strongly supported by this evidence. Second, we consider recent attempts by Eisner and Nadiri [6] and by Eisner [7] to estimate the elasticity of substitution from data on investment expenditures. The resulting estimates are extremely sensitive to errors in specification and prove to be highly unreliable. We conclude that these estimates of the elasticity of substitution from data on investment do not provide a useful alternative to our hypothesis that the elasticity of substitution is unity.

In the original study of the elasticity of substitution by Arrow, Chenery, Minhas, and Solow [1, henceforward ACMS] the elasticity of substitution was significantly different from zero for every industry group included in the study. For nine of the twenty-four industry groups the elasticity of substitution was significantly below unity. These findings are shown in the first column of Table 1. Fuchs [8] has re-examined these conclusions, inserting a dummy variable representing the level of development of countries included in the sample. For only two industry groups is the elasticity of substitution significantly different from unity—for one group the elasticity is above unity and for the other below. The results of ACMS and Fuchs are compared in Table 1. The first conclusion supported by these results is that the elasticity of substitution is significantly different from zero. However, Fuchs has thrown into question the finding of ACMS that the elasticity of substitution may be less than unity.

The results of ACMS and Fuchs are based on international cross sections. For the United States a number of studies have been made of the elasticity of substitution for two-digit industries within manufacturing. Dhrymes [5], Minasian [20], and Solow [23] have fitted relationships involving value added per man and the wage rate. More recently, this relationship has been fitted to data for successive cross sections by Griliches [9] and by Zarembka [26]. Second, Bell [2] and Dhrymes [5] have fitted analogous relationships involving capital and its rental price. All of these estimates provide support

TABLE 1—COMPARISON OF ESTIMATES OF THE ELASTICITY OF SUBSTITUTION FOR MANUFACTURING INDUSTRIES BY ACMS AND BY FUCHS FROM INTERNATIONAL CROSS SECTIONS, 1950–1955*

Industry Group	ACMS	Fuchs
Food and kindred products		
Dairy products	0.72	0.90
	(.07)	(.08)
Fruit and vegetable canning	0.86	1.09
	(.08)	(.10)
Grain and mill products	0.91	1.32
	(.10)	(.17)
Bakery products	0.90	1.07
	(.07)	(.11)
Sugar	0.78	0.90
	(.12)	(.18)
Tobacco	0.75	1.22
	(.15)	(.21)
Textile mill products		
Spinning and weaving	0.81	0.98
	(.07)	(.10)
Knitting mills	0.79	0.95
	(.06)	(.08)
Lumber and wood products	0.86	1.08
	(.07)	(.14)
Furniture and fixtures	0.89	1.04
	(.04)	(.09)
Pulp, paper, and products	0.97	0.91
	(.10)	(.18)
Printing and publishing	0.87	1.02
	(.06)	(.09)
Chemicals and products		
Basic chemicals	0.83	1.11
	(.07)	(.10)
Misc. chemicals	0.90	1.06
	(.06)	(.09)
Fats and Oils	0.84	1.06
	(.09)	(.18)
Leather and leather goods	0.86	0.98
	(.06)	(.10)
Stone, clay, and glass products		
Clay products	0.92	0.66
	(.10)	(.20)
Glass	1.00	1.27
	(.08)	(.10)
Ceramics	0.90	1.08
	(.04)	(.13)
Cement	0.92	1.31
	(.15)	(.22)
Primary metal products		
Iron and steel	0.81	0.76
	(.05)	(.11)
Nonferrous metals	1.01	0.94
	(.12)	(.20)
Fabricated metal products	0.90	1.01
	(.09)	(.17)
Electrical Machinery	0.87	1.03
	(.12)	(.21)

* Source: Nerlove [22], Table 1, Columns 1 and 2, pp. 60–63.

for the hypothesis that the elasticity of substitution is unity.

Minasian's estimates of the elasticity of substitution are based on data for individual states for 1957; he finds two estimates of the elasticity of substitution significantly above unity, three significantly below, and thirteen not significantly different from unity among a total of eighteen industry groups.[2] Solow's results are based on data for Census regions for 1956; he finds no estimates significantly below unity, only one significantly above, and seventeen not significantly different from unity.[3] Dhrymes employs data for individual states in 1957, essentially the same data analyzed by Minasian; however, he finds very low estimates of the elasticity of substitution for regressions in which value added per man is the dependent variable and the wage rate is an independent variable.[4] Zarembka has recently analyzed these same data and has discovered that Dhrymes inadvertently regressed the wage rate on value added per man rather than vice versa. With this error removed, estimates based on data for individual states in 1957 are similar to those of Minasian.[5]

In fitting relationships involving capital and its rental price to data for individual states for 1958, Bell finds no estimates significantly below unity, eleven significantly above, and seven not significantly different

[2] See [9], Table 1, p. 287.
[3] See [9], Table 1, p. 287.
[4] Dhrymes [5], Table 1, p. 358, column labeled σ_1.
[5] Zarembka [26], p. 9, fn. 2. Dhrymes' estimates are reproduced by Nerlove [22], Table 1, pp. 60–63, column labeled "Regression I." Corrected estimates are given by Zarembka, Table 1, p. 10, column labeled $\hat{\sigma}$ for 1957.

from unity.[6] Analyzing data on capital for 1957, Dhrymes finds one estimate significantly below unity, two significantly above, and fourteen not significantly different from unity.[7] The overall conclusion from these four studies of cross sections of individual states or Census regions for the years 1956, 1957, and 1958 is that the elasticity of substitution is not significantly different from unity. Significant deviations above unity are about as frequent as significant deviations below unity. We conclude that our maintained hypothesis that the production function is Cobb-Douglas in form is supported by evidence from studies of factor demand, including demand for labor and demand for capital.

Estimates of the elasticity of substitution from time series are much more erratic than those from cross sections. Estimates substantially above unity and substantially below unity have been obtained for alternative specifications of factor demand equations. The estimates appear to be extremely sensitive to changes in the stochastic specification of factor demand relationships; minor differences in measurement produce substantially different estimates.[8] The primary explanation for the lack of reliability of time series estimates of the elasticity of substitution is that the elasticity is essentially a second-order parameter in the relationship between the growth of output and the growth of factor inputs, as Nelson [21] has demonstrated.

A partial reconciliation of time series and cross section results has been undertaken by Griliches [9] and by Zarembka [26], taking into account variations in labor quality over the cross section of individual states and allowing for serial correlation in successive cross sections. Griliches estimates the elasticity of substitution for two-digit industries from cross sections of individual state observations for 1957. The results are given in the first column of Table 2. Griliches summarizes these results as follows:

[6] See [9], Table 1, p. 287.

[7] Dhrymes [5], Table 1, p. 358, column labeled σ_2.

[8] Time series studies are reviewed by Nerlove [22], pp. 82–100, and by Griliches [9], pp. 285–290.

Table 2—Estimates of the Elasticity of Substitution for Two-Digit Manufacturing Industries in the United States by Griliches*

Industry Group	a. Cross Section	b. Lagged Adjustment	c. Serial Correlation
Food	0.91 (.10)	1.01	**
Textile	0.94 (.17)	1.11	1.09 (.54)
Apparel	1.06 (.19)	0.84	0.63 (.21)
Lumber	1.07 (.06)	1.11	1.18 (.20)
Furniture	1.04 (.07)	0.99	**
Paper	1.67 (.30)	1.30	**
Printing	0.83 (.18)	0.45	0.68 (.28)
Chemicals	0.71 (.22)	0.59	0.70 (.40)
Petroleum	***	***	***
Rubber and Plastics	1.28 (.42)	2.21	0.90 (.43)
Leather	0.84 (.26)	1.60	1.16 (.40)
Stone, clay, glass	0.91 (.19)	0.77	1.88 (.49)
Primary metals	1.41 (.42)	3.49	2.37 (.47)
Fabricated metals	0.85 (.14)	1.17	1.20 (.28)
Machinery, exc. electrical	1.24 (.38)	2.40	2.00 (.29)
Electrical machinery	0.66 (.31)	0.40	1.53 (.53)
Transportation equipment	0.96 (.55)	1.09	**
Instruments	0.75 (.43)	0.82	**

* Source: Griliches [9], Table 3, p. 293.

** Contradicts serial correlation model; see Griliches [9], note c, Table 3, p. 294.

*** No significant relationships found.

> The first set of σ (elasticity of substitution) estimates is comparable to, though substantially better than (in terms of fit and t ratios) the Minasian and Solow estimates and is generally of the same order of magnitude. Only one of these σ's (out of 17) is significantly different from unity, and that one is above unity.[9]

[9] [9], p. 292.

Griliches next fits two alternative dynamic models to cross sections of individual states for 1957 and 1958—a lagged adjustment model and a model that takes serial correlation into account. The results are given in the second and third columns of Table 2. Griliches summarizes these results as follows:

> The second set of σ (elasticity of substitution) estimates is based on the partial adjustment equation, while the third is based on the serial correlation model. In 12 out of 17 cases the latter model is the one consistent with the data. In general, all of the estimated σ's are not very (statistically) different from unity, the significant deviations if anything occurring above unity rather than below it.[10]

In Griliches' serial correlation model equality between elasticities of substitution for cross sections in 1957 and 1958 is taken as a maintained hypothesis. This hypothesis can be tested by estimating elasticities of substitution separately for the two cross sections and testing their equality, given serial correlation in the residuals. A test of this type has been carried out by Zarembka. Except for one industry, the hypothesis is accepted. Zarembka also presents combined estimates of the elasticity of substitution, using a statistical model similar to that of Griliches. For only two industries of the thirteen studied by Zarembka is the elasticity of substitution significantly different from unity; for both the elasticity is significantly less than unity.[11] Our overall conclusion for estimates of the elasticity of substitition from successive cross sections is that this parameter is not significantly different from unity. For the individual cross sections analyzed by Bell, Dhrymes, Griliches, Minasian, Solow, and Zarembka, the elasticity of substitution is significantly different from zero. This conclusion also holds for the estimates from successive cross sections by Griliches and Zarembka.

We have already indicated that estimates of the elasticity of substitution from time series data on capital input are unreliable. This implication is strongly supported by the estimates of the elasticity of substitution from quarterly data on investment expenditures reported by Eisner [7]. Eisner's empirical results extend those of Eisner and Nadiri [6], incorporating a variable that represents changes in the difference between two measures of the rental price of capital. Eisner demonstrates that the effect of this variable is not significantly different from zero in any of the regressions he presents so that Eisner's model effectively reduces to that presented by Eisner and Nadiri.

The Eisner-Nadiri model is alleged to provide a direct test of a model presented earlier by Jorgenson and Stephenson [14] in which the elasticity of substitution is taken equal to unity. Closer examination of the two models reveals that they are mutually contradictory and that evidence predicated on the validity of one is not relevant to the validity of the other.[12] In addition, Bischoff [3] has drawn attention to an important error of specification in the econometric model used by Eisner and Nadiri. Since the model presented by Eisner effectively reduces to that of Eisner and Nadiri, Bischoff's analysis carries over directly to Eisner's model. The basic difficulty in the Eisner-Nadiri model arises from the fact that errors in the distributed lag function employed in the model are autocorrelated. Eisner and Nadiri have ignored this autocorrelation so that their estimates of parameters of the distributed lag function are inconsistent. The estimated elasticity of substitution is highly sensitive to this error in specification. Bischoff demonstrates that elimination of the error in specification from the model used by Eisner and Nadiri and by Eisner results in the following conclusions:

1. The elasticity of capital with respect to the ratio of the price of output to the rental price of capital is not significantly different from unity.[13]

2. The elasticities of capital with respect

[10] [9], p. 292.

[11] Zarembka [26], Table 2, p. 16.

[12] Jorgenson and Stephenson [17], Section 2, pp. 3–5.

[13] Bischoff [3], Table 2, tests for the maintained hypothesis that the errors are a first order autoregressive process. This conclusion directly contradicts the first conclusion of Eisner and Nadiri [6], p. 380.

to output and prices taken together are not significantly different from unity.[14]

3. The best fitting elasticity of capital with respect to output is negative with a high standard error, hardly consistent with flexible accelerator models of investment as Eisner and Nadiri and Eisner have claimed.[15]

4. Results do not contradict our maintained hypothesis that the production function is Cobb-Douglas in form.[16]

Our conclusion is that the results of Eisner [7] and Eisner and Nadiri [6] rest on an error in the stochastic specification of the Eisner-Nadiri model and that all of the conclusions Eisner has drawn concerning our analysis of tax policy are directly reversed for a correctly specified model of the Eisner-Nadiri variety. The sharp conflict between estimates of the elasticity of substitution by Eisner and Nadiri and estimates from individual and successive cross sections we have reviewed above is removed when the error in specification of the Eisner-Nadiri model is eliminated. Estimates of parameters such as the elasticity of substitution and the degree of homogeneity of the production function from the Eisner-Nadiri model are highly unreliable. This fact is demonstrated both by the large standard errors associated with estimates from a correctly specified model and by the high degree of sensitivity of the estimates to errors in specification. These estimates do not provide a useful alternative to the hypothesis that the production function is Cobb-Douglas in form.

We turn now to Coen's estimates of the elasticity of substitution. His model is a straightforward generalization of ours, obtained by replacing our demand function for capital by one derived from a *CES* production function. Thus he is able to avoid the specification error of Eisner and Nadiri, but at the cost of estimating a nonlinear equation. His estimates of σ are clearly inconsistent with the empirical results on the elasticity of substitution that we have summarized above. Unfortunately, Coen does not test the hypothesis that σ is different from unity, so the reader has no information about the statistical significance of his results. But whether or not his estimates are significantly different from one, it is not appropriate to proceed on the basis that σ is in the neighborhood of 0.2 in investment research in the face of the large body of evidence to the contrary from research on production functions and from Bischoff's work on investment functions. Until a reconciliation is available for the low elasticities of substitution in Coen's investment equations and the roughly unitary elasticities in production equations, the most reasonable course is to impose the value of unity in investment equations.

II. *The Conceptual Basis for the Measurement of the Effects of Tax Policy.*

Coen expresses concern about the "legitimacy" of the appearance of output in the demand function for capital; he concludes that investigators who have studied investment equations in which an output variable appears in the right-hand side have made a serious conceptual error. Thus Coen rejects our demand function for capital,

$$K_t^* = \alpha \frac{p_t Q_t}{c_t}$$

on the grounds that Q_t is set by the firm and cannot "legitimately" appear on the right hand side. What he presumably means is that this equation could not be used for decision-making within the firm; until the firm decides on Q_t, it cannot calculate K_t^*.[17]

Our interpretation of this equation is the following: It is the demand for capital conditional on the level of output. The question

[14] Bischoff [3], Table 2. This conclusion directly contradicts the second conclusion of Eisner and Nadiri [6], p. 380.

[15] The numerical value given by Bischoff is $-.434$ under the hypothesis that the autoregressive parameter is zero. This contradicts the third conclusion of Eisner and Nadiri [6], pp. 380–81. This conclusion is repeated by Eisner [7], pp. 5–6.

[16] This contradicts the fourth conclusion of Eisner and Nadiri [6], p. 381. The remaining conclusions of Eisner and Nadiri are refuted by Bischoff [3] and by Jorgenson and Stephenson [17], esp. Section 4, pp. 11–12.

[17] Coen's argument must be distinguished from the entirely different econometric point that if a jointly dependent variable appears on the right-hand side the choice of estimation method should take account of that fact.

of how firms determine output is an interesting one, but, as we shall see, it is irrelevant to our particular use of the investment equation. We should note a very serious difficulty which Coen's approach might encounter: if firms produce under constant returns to scale, their choice of output is indeterminate. Our conditional demand function still holds, but Coen's equation breaks down. The reader is invited to substitute parameter values with $\alpha+\beta$ equal to one into Coen's equation 17 to see how this difficulty arises. Since empirical evidence generally supports constant returns to scale,[18] Coen's method has serious shortcomings.

The point of Coen's argument is that tax policy affects not only the conditional demand for capital given output, but also the level of output itself. Since we measured only the first effect and assumed that output remained unchanged, Coen suggests that we understated the true effect of tax policy. Coen's charge that we thereby committed a conceptual error cannot be sustained. At worst, we failed to make clear the question we were answering with our calculations. To be precise, the question was the following: Suppose that instead of increasing investment incentives, the government had pursued an alternative fiscal policy which resulted in exactly the same dollar value of output, interest rate, and capital goods price. Then by how much would investment have been less under the alternative policy?

Another question which could be asked is the general equilibrium one: What were the overall effects of tax incentive policies, assuming no alternative compensatory policy? The answer waits for the development of a satisfactory model of general equilibrium in the United States economy.

Coen prefers to ask a third question: Suppose that there had been no investment incentive policy and that the government had pursued an alternative fiscal policy that reduced output to the point that every firm was in equilibrium at the output price, wage, interest rate, and capital goods price that existed under the incentive policy. Then by how much would investment have been reduced? The difficulty with this question is that unless decreasing returns to scale are strong, a very large decrease in output is necessary to compensate for the higher rental price of capital goods in the absence of the investment incentives. In the case of constant returns to scale, output and capital would always fall to zero in this circumstance. This suggests that Coen's question is rather less interesting than he supposes. By contrast, our calculations are not affected by the assumption of constant returns to scale. Even if a general equilibrium calculation could be made, we would want to carry out our calculations in order to separate the total effect of tax policy into an incentive effect and an output effect. We have attempted to provide a method for measuring the incentive effect which will be useful in further research in this area.

III. *New Estimates of the Parameters of Investment Functions.*

Our earlier evaluation of the effects of tax policy on investment behavior was based on policy changes through 1963. Since 1963, a number of important changes in tax policy have taken place, providing a further opportunity for evaluation of the responsiveness of investment expenditures to changes in tax policy. For this purpose we have re-estimated our model, using data that have become available since our earlier study. We have also revised our econometric technique to take account of recently developed methods for analysis of distributed lag functions. With these changes we obtain a new set of investment functions for the non-farm sector of the United States. In the following section we employ these investment functions to evaluate the effects of changes in tax policy that have taken place since 1963.

Our basic econometric model is a dis-

[18] Zarembka has examined this question very carefully in the study previously cited. For three of his thirteen industry groups the degree of homogeneity differs significantly from unity; for one of these groups, the degree of homogeneity is below unity as assumed by Coen. For the other two, the degree of homogeneity is above unity. Overall, there is little evidence that the degree of homogeneity differs significantly from unity; deviations above unity appear, if anything, to be slightly more important than deviations below unity.

tributed lag function in net investment and changes in desired capital:

$$N_t = \sum_{\tau=0}^{4} \beta_\tau (K^*_{t-\tau} - K^*_{t-\tau-1}) + \epsilon_t$$

where N_t is net investment in period t, K^*_t is desired capital, and ϵ_t is a random error.[19] Under the assumption that the production function is Cobb-Douglas in form, the desired level of capital is given by:

$$K^*_t = \alpha \frac{p_t Q_t}{c_t},$$

where α is the elasticity of output with respect to capital, p_t is the price of output, Q_t the quantity, and c_t the rental price of capital.

The rental price of capital services depends on the tax rate u, the after tax rate of of return r, the investment goods price q, the rate of replacement δ:

$$c = \frac{1 - k - uz}{1 - u} q(r + \delta),$$

where z is the discounted value of depreciation allowances allowed for tax purposes and k is the investment tax credit. This formula is appropriate for the period after 1964 when the tax credit is not deducted from allowable depreciation. For 1962 and 1963 the appropriate formula is:

$$c = \frac{(1 - k)(1 - uz)}{1 - u} q(r + \delta);$$

during these years the depreciation base was reduced by the amount of the investment tax credit.

Through 1953 the rental price is that appropriate to straight line depreciation. Since 1954 the rental price is that appropriate to the sum of the years' digits depreciation.[20] The investment tax credit was introduced in 1962 at a rate nominally equal to 7 percent of the value of investment in equipment. In practice, certain limitations on the applicability of the investment tax credit reduce its effective rate to 6 percent for manufacturing equipment and 5.8 percent for non-farm, non-manufacturing equipment.[21] From October 1966 to March 1967 the investment tax credit was suspended.

We took the tax rate to be the statutory rate prevailing during most of each year. We did not allow for excess profits taxes during the middle thirties or the Korean War. For all years we took the rate of return before taxes ρ to be constant at 20 percent. This value is higher than the value of 14 percent used in our previous studies. The higher value is consistent with the results of Jorgenson and Griliches [15]. Under our assumption of a constant before-tax rate of return the after-tax rate $r = (1-u)\rho$ varies with the tax rate. The investment goods price is the same as that used to deflate investment expenditures in current prices and the rate of replacement is the same as that used to calculate capital stock. Estimates of lifetimes of assets allowable for tax purposes were obtained from a special Treasury study [24]. These estimates are the same as those employed in our previous studies:

Period	Equipment	Structures
1929–54	17.5	27.8
1955	16.3	25.3
1956–61	15.1	22.8
1962–65	13.1	22.8

New estimates of these lifetimes for recent years would require that the special Treasury study be updated.

The statistical technique for fitting our econometric model to data on investment expenditures is described in detail in our paper, [12]. Briefly, we include a lagged value of net investment together with current and five lagged values of the change in the desired level of capital. The coefficient of lagged net investment in this first stage regression is interpreted as an estimate of autocorrelation. It is used to perform an

[19] For further details, our paper [10], esp. pp. 392–98, may be consulted.

[20] Depreciation under the sum of the years' digits formula has a higher present value for the range of lifetimes and rates of return of interest for this study. See [10], Table 1, p. 395.

[21] These estimates of the effective rate of the tax credit are based on data from tax returns for 1963 [25].

TABLE 3—FITTED INVESTMENT FUNCTIONS, 1935–40 AND 1954–65, SECOND STAGE RESULTS

Asset Class	$\hat{\alpha}\hat{\beta}_0$	$\hat{\alpha}\hat{\beta}_1$	$\hat{\alpha}\hat{\beta}_2$	$\hat{\alpha}\hat{\beta}_3$	$\hat{\alpha}\hat{\beta}_4$	R^2	s	d
Mfg. Equipment	.0130 (.0047)	.0200 (.0034)	.0208 (.0040)	.0153 (.0040)	.0036 (.0053)	.602	.620	2.099
Mfg. Structures	.0041 (.0033)	.0082 (.0030)	.0093 (.0035)	.0073 (.0032)	.0024 (.0034)	.186	.513	1.304
Non-farm, Non-mfg. Equipment	.0374 (.0083)	.0282 (.0038)	.0211 (.0063)	.0160 (.0047)	.0129 (.0090)	.800	1.190	1.724
Non-farm, Non-mfg. Structures	.0059 (.0034)	.0105 (.0032)	.0118 (.0035)	.0098 (.0031)	.0046 (.0029)	.169	.506	1.825

autoregressive transformation of net investment and change in desired capital, which are then used in a second stage regression. In the second stage we impose the constraint that the lag coefficients lie on a parabola. The estimated lag functions and other regression results are given in Table 3. It should be noted that R^2 for these regressions is a measure of the degree of explanation of the autoregressively transformed values of net investment, and is therefore not comparable to the R^2 in our earlier results. The overall goodness of fit as measured by the standard errors is superior to that of our previous investment functions for 1931–41 and 1950–63 except for manufacturing structures. This improvement is mainly due to the change in time period and to revisions of the basic investment data; however, it is also partly due to the change in our specification of the distributed lag function. We have added three lagged changes in desired capital, which improves the results to some extent.

The parameters of the distributed lag function $\{\mu_i\}$ may be estimated by employing the constraint that the sum of the coefficients of this function must be unity to estimate the parameter α.[22] The resulting estimates are given in Table 4. The mean lag for each function is also given in Table 4. Comparing these mean lags with estimates from our earlier studies, we find that the new estimates were very similar for investment in equipment. The mean lag is now estimated to be slightly lower for manufacturing equipment and slightly higher for non-farm, non-manufacturing equipment. For structures, the new estimates differ substantially from the old. The old estimate of the mean lag for manufacturing structures was 3.84 years whereas the new estimate is 1.86; the old estimate of the mean lag for non-farm, non-manufacturing structures was 7.49 years whereas the new estimate is 1.92. For both sets of results the lags for structures are estimated to be longer than for equipment.

A disturbing feature of our earlier results is that the lag pattern fails to agree with the substantial body of evidence from studies at

TABLE 4—FITTED INVESTMENT FUNCTIONS, 1935–40 AND 1954–65, DERIVED RESULTS

Asset Class	Mean Lag (years)	$\hat{\alpha}$
Manufacturing Equipment	1.67	.0727
Manufacturing Structures	1.86	.0312
Non-farm, Non-manufacturing Equipment	1.47	.1160
Non-farm, Non-manufacturing Structures	1.92	.0426

$\hat{\alpha}$: Estimate of the elasticity of output with respect to the capital input.

[22] For detailed discussion of this restriction and its use in estimation of the parameter α, see Jorgenson [13], pp. 135, 147–48.

the level of two-digit industries by Jorgenson and Stephenson [14] and studies at the level of the individual firm by Jorgenson and Siebert [16]. For manufacturing, Jorgenson and Stephenson estimate the average lag at about two years, while results from individual industries range from six to eleven quarters with results clustering in the neighborhood of the overall average. The results for individual firms are characterized by more variability than the results for industries, as would be expected. The average lags estimated by Jorgenson and Siebert range from less than a year to over three years with values between one and two years predominating. "Mayer estimated a seven quarter average lag from the decision to undertake investment to the completion of the project for manufacturing and electric power combined on the basis of surveys [19]. We conclude that our new estimates agree closely with Mayer's survey results and with estimates derived from investment functions for industry groups and for individual firms. Our previous estimates of the average lags for structures are evidently biased by specification errors in the underlying distributed lag functions and should be replaced by our new estimates.

IV. *The Impact of Tax Policy on Investment Behavior.*

Two major changes in tax policy affecting investment behavior have been made since 1963. First, in the tax cut of 1964, the corporate income tax rate was reduced from 52 percent to 48 percent. At the same time, business firms were no longer required to reduce the base for depreciation by the amount of the investment tax credit. This alteration in the tax law substantially enhanced the effectiveness of the investment tax credit adopted in 1962. A second change in tax policy is that from October 1966 to March 1967 the investment tax credit and certain types of accelerated depreciation were suspended. In the legislation implementing suspension of the tax credit, the period of suspension was to run from October 1966 to December 1967. We analyze the effects of the suspension that actually took place and the hypothetical effects of a longer suspension, as stipulated at the time the suspension began.

The qualitative features of the response of investment to a change in tax policy are essentially the same for all changes. To evaluate the effects of particular tax measures it is useful to assess the response of investment quantitatively. Our calculations are based on a partial equilibrium analysis of investment behavior. We hold all determinants of investment expenditures except for tax policy equal to their actual values. We then measure the impact of tax policy by substituting into our investment functions parameters of the tax structure—tax rate, depreciation formulas, and tax credit—appropriate to alternative tax policies. The difference between investment resulting from actual tax policy and investment that would have resulted from alternative tax policies is our measure of the impact of tax policy. In our new calculations both investment and capital stock are measured in prices of 1965. We estimate the impact of all changes in tax policy through 1970. In order to make these estimates, we employed a rough set of projections of the determinants of investment. No great precision was required in these projections, since the estimates of the differential impacts of alternative policies are not at all sensitive to the assumed level of investment.[23]

In analyzing the effect of the tax cut, we assume that the before tax rate of return was left unchanged. Under this condition, the effect of a change in the tax rate on the rental price of capital services is neutral, provided that depreciation for tax purposes is equal to economic depreciation.[24] Under the conditions actually prevailing in 1964, depreciation for tax purposes was in excess of economic depreciation for both plant and equipment in manufacturing and non-farm, non-manufacturing sectors. Accordingly, the rental price of capital services resulting from the tax cut was actually greater than the rental price before the cut. Following are the

[23] Further details are given in [12], pp. 54–55.

[24] See Section 2 of our paper [12].

TABLE 5. CHANGES IN GROSS INVESTMENT (I), NET INVESTMENT (N), AND CAPITAL STOCK (K) RESULTING FROM THE TAX CUT OF 1964 (Billions of 1965 Dollars)

Year	Manufacturing						Non-Farm, Non-Manufacturing					
	Equipment			Structures			Equipment			Structures		
	I	N	K	I	N	K	I	N	K	I	N	K
1964	−.049	−.049	0	−.020	−.020	0	−.135	−.135	0	−.044	−.044	0
1965	−.136	−.129	−.049	−.062	−.061	−.020	−.267	−.241	−.135	−.125	−.123	−.044
1966	−.181	−.155	−.178	−.088	−.083	−.081	−.217	−.145	−.376	−.167	−.155	−.167
1967	−.186	−.137	−.333	−.087	−.077	−.164	−.223	−.123	−.521	−.168	−.146	−.322
1968	−.155	−.086	−.470	−.066	−.051	−.241	−.274	−.150	−.644	−.142	−.110	−.468
1969	−.120	−.038	−.556	−.043	−.025	−.292	−.245	−.092	−.794	−.099	−.059	−.578
1970	−.119	−.032	−.594	−.043	−.023	−.317	−.213	−.043	−.886	−.088	−.044	−.637

rental prices for 1965, the first full year of the tax cut:

	Without Tax Cut	With Tax Cut
Manufacturing Equipment	.296	.299
Manufacturing Structures	.237	.240
Non-Farm, Non-Manufacturing Equipment	.352	.355
Non-Farm, Non-Manufacturing Structures	.247	.250

Our estimates of the decrease in net investment, gross investment, and capital stock resulting from this change are given in Table 5. In general, the effects of the rate reduction are small and negative. It should be emphasized that these estimates are conditional on the level of output actually resulting from the tax cut; quite clearly the overall effect of the tax cut was to stimulate investment by increasing output. A second, little noticed change in tax policy in 1964 was the repeal of the Long Amendment; after repeal, the tax credit was no longer deducted from the depreciation base for tax purposes. This change raises the effective rate of the tax credit to almost 10 percent as compared with approximately 6 percent under the Long Amendment. The rental price of capital services for equipment in 1964 with and without repeal of the Long Amendment are:

	With Long Amendment	Without Long Amendment
Manufacturing Equipment	.302	.293
Non-Farm, Non-Manufacturing Equipment	.363	.352

Estimates of the increase in net investment, gross investment, and capital stock resulting from this change are given in Table 6.

These increases are quite substantial. The peak effect for manufacturing equipment took place in 1965 at which time the net investment in equipment attributable to the repeal was 10.4 percent of total net investment. In the non-farm, non-manufacturing sector, the peak effect in 1964 was over a billion dollars and accounted for 16.3 percent of net investment in equipment in that sector. Once again, a dip in the effect of this policy change can be seen in 1966–67 and one or two years after, resulting from the suspension of the investment credit. The lag structure in the non-manufacturing sector makes the dip much more noticeable there than in the manufacturing sector.

TABLE 6—CHANGES IN GROSS INVESTMENT (I), NET INVESTMENT (N), and CAPITAL STOCK (K) RESULTING FROM THE REPEAL OF THE LONG AMENDMENT (Billions of 1965 Dollars)

Year	Manufacturing Equipment			Non-Farm, Non-Manufacturing Equipment		
	I	N	K	I	N	K
1964	.238	.238	0	1.042	1.042	0
1965	.400	.365	.238	.958	.758	1.042
1966	.412	.329	.603	.706	.360	1.800
1967	.349	.217	.932	.750	.335	2.160
1968	.229	.067	1.149	1.021	.541	2.495
1969	.236	.064	1.216	.761	.177	3.036
1970	.297	.115	1.280	.792	.174	3.213

TABLE 7—CHANGE IN GROSS INVESTMENT (I), NET INVESTMENT (N), AND CAPITAL STOCK (K) RESULTING FROM SUSPENSION OF THE INVESTMENT TAX CREDIT FOR EQUIPMENT AND ACCELERATED DEPRECIATION FOR STRUCTURES FROM OCTOBER 10, 1966 TO MARCH 8, 1967 (Billions of 1965 Dollars)

Year	Manufacturing						Non-Farm, Non-Manufacturing					
	Equipment			Structures			Equipment			Structures		
	I	N	K	I	N	K	I	N	K	I	N	K
1966	−.177	−.177	0	−.046	−.046	0	−.762	−.762	0	−.119	−.119	0
1967	−.271	−.245	−.177	−.089	−.086	−.046	−.599	−.452	−.762	−.200	−.192	−.119
1968	−.153	−.091	−.422	−.060	−.052	−.132	.069	.303	−1.214	−.126	−.104	−.311
1969	−.009	.066	−.513	.011	.012	−.184	.051	.226	−.911	−.011	.018	−.415
1970	.157	.223	−.447	.074	.075	−.172	.018	.150	−.685	.111	.139	−.397

TABLE 8—CHANGE IN GROSS INVESTMENT (I), NET INVESTMENT (N), AND CAPITAL STOCK (K) RESULTING FROM HYPOTHETICAL SUSPENSION OF THE INVESTMENT TAX CREDIT FOR EQUIPMENT AND ACCELERATED DEPRECIATION FOR STRUCTURES FROM OCTOBER 10, 1966 TO DECEMBER 31, 1967 (Billions of 1965 Dollars)

Year	Manufacturing						Non-Farm, Non-Manufacturing					
	Equipment			Structures			Equipment			Structures		
	I	N	K	I	N	K	I	N	K	I	N	K
1966	−.177	−.177	0	−.046	−.046	0	−.762	−.762	0	−.119	−.119	0
1967	−.872	−.846	−.177	−.250	−.247	−.046	−3.190	−3.043	−.762	−.614	−.606	−.119
1968	−.567	−.416	−1.023	−.234	−.216	−.293	.208	.940	−3.805	−.475	−.425	−.725
1969	−.181	.031	−1.439	−.063	−.032	−.509	.171	.722	−2.865	−.154	−.074	−1.150
1970	.270	.477	−1.408	.117	.152	−.541	.093	.505	−2.143	.193	.278	−1.224

In 1966, an important objective of economic policy was to restrain investment. After a number of alternative changes in tax policy were considered and rejected,[25] the investment tax credit for equipment was suspended beginning October 10, 1966; at the same time, accelerated depreciation for structures was replaced by 150 percent declining balance depreciation. In the original legislation implementing these changes in tax policy, the suspension was to remain in effect until the end of 1967, a total period of almost fifteen months. The suspension was lifted on March 9, 1967, so that the total period of suspension was a little less than five months. The effect of the suspension on the annual rental price of capital in 1967 was the following:

	Without Suspension	With Suspension
Manufacturing Equipment	.320	.351
Manufacturing Structures	.259	.276
Non-Farm, Non-Manufacturing Equipment	.379	.414
Non-Farm, Non-Manufacturing Structures	.270	.287

Our estimates of the effects of the suspension on net investment, gross investment, and capital stock are given in Table 7.

For all categories of assets, the suspension had a restraining effect on the level of investment in 1967. We estimate that this effect continued into 1968 for all assets except non-farm, non-manufacturing equipment. For all classes of assets, the restoration of the original tax credit for equipment and accelerated depreciation for structures will result in a stimulus to investment in 1969 and 1970. For no class of assets is the level of capital stock as high at the end of 1970 as it would

[25] Policies under consideration during early 1966 and their potential impact on investment expenditures are discussed in our earlier study [11].

have been in the absence of the suspension. The total gross investment for the five year period 1966–70 is considerably lower than it would have been in the absence of the five month suspension.

If the suspension of the investment tax credit for equipment and accelerated depreciation for structures had continued for fifteen months, the impact on the level of investment would have been much more substantial. Our estimates are given in Table 8. For investment in structures the restraining effect of the suspension would have continued into 1969, although the impact would have been very slight in that year. For investment in equipment as well as for structures the magnitude of the impact would have been much greater. As a result the stimulus from restoration of the tax credit and accelerated depreciation would have been correspondingly increased.

References

1. K. J. Arrow, H. B. Chenery, B. S. Minhas, and R. M. Solow, "Capital-Labor Substitution and Economic Efficiency," *Rev. Econ. Stat.*, Aug. 1961, *43*, 225–50.
2. F. W. Bell, "The Role of Capital-Labor Substitution in the Economic Adjustment of an Industry Across Regions," *So. Ec. Jour.*, Oct. 1964, *31*, 123–31.
3. C. W. Bischoff, "Hypothesis Testing and the Demand for Capital Goods," *Rev. Econ. Stat.*, Aug. 1969, *51*, forthcoming.
4. R. M. Coen, "Tax Policy and Investment Behavior: Comment," *Am. Econ. Rev.*, June 1969, *60*, 370–79.
5. P. Dhrymes, "Some Extensions and Tests for the CES Class of Production Functions," *Rev. Econ. Stat.*, Nov. 1965, *47*, 357–66.
6. R. Eisner and M. I. Nadiri, "Investment Behavior and Neo-Classical Theory," *Rev. Econ. Stat.*, Aug. 1968, *50*, 369–82.
7. ———, "Tax Policy and Investment Behavior: Comment," *Am. Econ. Rev.*, June 1969, *60*, 379–88.
8. V. R. Fuchs, "Capital-Labor Substitution: A Note," *Rev. Econ. Stat.*, Nov. 1963, *45*, 436–38.
9. Z. Griliches, "Production Functions in Manufacturing: Some Preliminary Results," in M. Brown (ed.), *The Theory and Empirial Analysis of Production*, Studies in Income and Wealth, No. 31. New York: *National Bureau of Economic Research*, New York 1967, 275–322.
10. R. E. Hall and D. W. Jorgenson, "Tax Policy and Investment Behavior," *Am. Econ. Rev.*, June 1967, *57*, 391–414.
11. ——— and ———, "The Role of Taxation in Stabilizing Private Investment," in V. P. Rock (ed.), *Policy Makers and Model Builders*, forthcoming.
12. ——— and ———, "The Quantitative Impact of Tax Policy on Investment Expenditures," in Gary Fromm (ed.), *Effect of Investment Tax Credit and Accelerated Depreciation on the Level of Investment*, Washington, D.C., The Brookings Institution, forthcoming.
13. D. W. Jorgenson, "Rational Distributed Lag Functions," *Econometrica*, Jan. 1966, *34*, 135–49.
14. ———, and J. A. Stephenson, "The Time Structure of Investment Behavior in U. S. Manufacturing, 1947–1960," *Rev. Econ. Stat.*, Feb. 1967, *49*, 16–27.
15. ———, and Z. Griliches, "The Explanation of Productivity Change," *Rev. Econ. Stud.*, July 1967, *34*, 249–83.
16. ———, and C. D. Siebert, "Optimal Capital Accumulation and Corporate Investment Behavior," *Jour. Pol. Econ.*, Nov./Dec. 1968, *76*, 1123–51.
17. ———, and J. A. Stephenson, "Issues in the Development of the Neo-Classical Theory of Investment Behavior," *Rev. Econ. Stat.*, Aug. 1969, *51*, forthcoming.
18. J. Kmenta, "On Estimation of the CES Production Function," *Internat. Econ. Rev.*, June 1967, *8*, 180–89.
19. T. Mayer, "Plant and Equipment Lead Times," *Jour. Bus.*, Apr. 1960, *33*, 127–32.
20. J. R. Minasian, "Elasticities of Substitution and Constant-Output Demand

Curves for Labor," *Jour. Pol. Econ.*, June 1961, *69*, 261–70.
21. R. R. Nelson, "Aggregate Production Functions and Medium-Range Growth Projections," *Am. Econ. Rev.*, Sept. 1964, *54*, 575–606.
22. M. Nerlove, "Recent Empirical Studies of the CES and Related Production Functions," in M. Brown (ed.), *The Theory and Empirical Analysis of Production*, Studies in Income and Wealth, New York: *National Bureau of Economic Research*, New York 1967, 55–122.
23. R. M. Solow, "Capital, Labor and Income in Manufacturing," in Conference on Research in Income and Wealth, *The Behavior of Income Shares: Selected Theoretical and Empirical Issues*, Studies in Income and Wealth, No. 27, Princeton, 1964, 101–28.
24. U. S. Department of Treasury, Bureau of Internal Revenue, "Supplementary Depreciation Data from Corporate Returns," *Statistics of Income*, 1959.
25. ———, "Supplementary Depreciation Data from Corporate Returns," *Statistics of Income*, 1963.
26. P. Zarembka, "On the Empirical Relevance of the CES Production Function," Inst. Inter. Studies, Univ. Calif., Berkeley, Tech. Report No. 9, Jan. 1969, submitted for publication.

ALTERNATIVE INTEREST RATES AND THE DEMAND FOR MONEY: COMMENT

By Harvey Galper*

In a recent issue of this *Review*, T. H. Lee [3] examined the influence on the demand for money of the interest rate paid on savings and loan (*S* and *L*) shares. His regressions revealed a strong effect of the *S* and *L* rate on holdings of money. Furthermore, the *S* and *L* rate had a stronger influence than all other rates examined, and, on an annual basis at least, complete adjustment of the actual to the desired money stock was found when the *S* and *L* rate was used. These findings imply strong substitution effects between holdings of money and savings and loan shares and call into question Friedman's [5] notion of the money stock which assumes that time deposits at commercial banks are the closest substitute to money proper.

Lee's conclusions, however, are based on only fifteen annual observations, and, in several instances, he mentions that the lack of quarterly data prevents the development of a more thorough analysis.[1] By interpolating semiannual data from the Federal Home Loan Bank Board (*FHLBB*), I have been able to develop a quarterly series on savings and loan rates for the period 1956–66. This series has then been applied in regressions similar to Lee's to see if his results would be changed by the use of quarterly data. By and large his conclusions have been confirmed, although quarterly data allow a more precise investigation of the lag structure in adjusting actual to desired money holdings.

In the following sections of this note, I shall first present the technique for deriving the quarterly savings and loan rate, along with the resulting series, and then shall present the regression results based on the interpolated quarterly series.

I. *Derivation of a Quarterly Series on Interest Rates Paid on Savings and Loan Shares*[2]

For the six month periods ending in June and December of each year, 1956–65, the Federal Home Loan Bank Board has prepared data on the average interest rates paid by all insured savings and loan associations. These rates are the advertised rates of each institution weighted by the savings capital held by the institution at the end of

* The author is at the Board of Governors of the Federal Reserve System. He would like to thank Frank de Leeuw and Edward Gramlich for their helpful comments in the preparation of this note.

[1] [3, for example, p. 1177 and fn. 11, p. 1178].

[2] This section, in particular, owes much to discussions with Frank de Leeuw.

TABLE 1—INTEREST RATES PAID ON SAVINGS AND LOAN SHARES, SEMI-ANNUALLY

Semi-annual Period	S and L Rate
1956–I	3.07
1956–II	3.18
1957–I	3.31
1957–II	3.43
1958–I	3.50
1958–II	3.48
1959–I	3.54
1959–II	3.74
1960–I	3.98
1960–II	4.02
1961–I	4.03
1961–II	4.06
1962–I	4.19
1962–II	4.24
1963–I	4.28
1963–II	4.27
1964–I	4.29
1964–II	4.31
1965–I	4.32
1965–II	4.33
1966–I	4.38
1966–II	4.65

Source: Federal Home Loan Bank Board.

the period. For 1966, these rates were not released in the same format but could be calculated from Table 2 of the *FHLBB* publication of June 1967, *Dividend Rates and the Structure of Savings Accounts*.[3] The semi-annual series so derived is presented in Table 1.

A regression equation was then employed to interpolate quarterly these semiannual observations. The interpolation equation was derived from a postulated rate setting function on the part of savings and loan associations. In terms of a rate adjustment process, it is assumed that the S and L rate is changed in each quarter so as to eliminate a fraction, k, of the difference between the desired S and L rate, R_s^*, and the previous period's rate, R_{s-1}. Thus,

$$\Delta R_s = k(R_s^* - R_{s-1}). \tag{1}$$

The association's desire for stability in interest rates paid as well as the effect of aggregation across all institutions prevents the adjustment from being made completely within one quarter.

It is then hypothesized that the desired rate, R_s^*, is a function of R_m and R_{td}, where the former is the interest rate paid on mortgages, the dominant asset in the S and L's portfolio, and the latter is the rate paid by a competing financial institution, in this case commercial bank time deposits. Thus,

$$R_s^* = a_0 + a_1 R_m + a_2 R_{td}. \tag{2}$$

Furthermore, the adjustment to a change in the time deposit rate is assumed to be accelerated by a change in the maximum rate payable on time deposits under regulation Q. This last variable, ΔR_{max}, is therefore an *impact* variable; which does not affect the desired S and L rate but merely the speed with which it is attained.[4] Substituting (2) into (1) and adding ΔR_{max} yields the follow-

[3] In contrast to the earlier series, however, this publication uses January and July as the two reporting dates for the first and second half of the year.

[4] The maximum rate payable on time deposits is likely to be an argument in most functions explaining the time deposit rate. See, for example, Patric Hendershott [4]. Therefore, one may question the inclusion of the change in this maximum rate along with the time deposit rate itself in deriving the S and L rate. The reasons for so doing are as follows. The largest and most abrupt changes in R_{td} occur after maximum rates have been raised, since commercial banks are then able to offer the time deposit rates they wish given their own portfolio structures and the prevailing pattern of market interest rates. At these times, S and L's must respond to the changes in R_{td} more quickly than they would otherwise do if they are to retain their competitive advantage in interest rates paid. Furthermore, the announcement effects associated with changes in the maximum rate tend to elicit faster S and L reactions. Indeed, as equation (9) indicates, R_s normally moves quite slowly in response to R_{td}, and the inclusion of R_{max} has the effect of accelerating this response. It should be noted that the *normal* speed of adjustment of R_s in response to R_{td} is not significantly altered when ΔR_{max} is omitted from the equation.

ing quarterly equation:

$$\Delta R_s = ka_0 + ka_1 R_m + ka_2 R_{td} + a_3 \Delta R_{max} - kR_{s_{-1}} \tag{3}$$

or, solving for R_s,

$$R_s = ka_0 + ka_1 R_m + ka_2 R_{td} + a_3 \Delta R_{max} + (1 - k) R_{s_{-1}} \tag{4}$$

This equation cannot be estimated as it stands since only semiannual data for R_s are available.[5] It can, however, be put into a form capable of estimation, but first let us make the following substitutions:

$$\begin{aligned} \alpha_0 &= ka_0 \\ \alpha_1 &= ka_1 \\ \alpha_2 &= ka_2 \\ \alpha_3 &= a_3 \\ \beta &= (1 - k) \end{aligned}$$

Solving for $R_{s_{-1}}$ in terms of $R_{s_{-2}}$ and substituting into (4) yields

$$\begin{aligned} R_s = {} & \alpha_0 + \alpha_1 R_m + \alpha_2 R_{td} + \alpha_3 \Delta R_{max} \\ & + \beta\alpha_0 + \beta\alpha_1 R_{m_{-1}} + \beta\alpha_2 R_{td_{-1}} \\ & + \beta\alpha_3 \Delta R_{max_{-1}} + \beta^2 R_{s_{-2}}. \end{aligned} \tag{5}$$

The corresponding equation for the lagged S and L rate is:

$$\begin{aligned} R_{s_{-1}} = {} & \alpha_0 + \alpha_1 R_{m_{-1}} + \alpha_2 R_{td_{-1}} \\ & + \alpha_3 \Delta R_{max_{-1}} + \beta\alpha_0 + \beta\alpha_1 R_{m_{-2}} \\ & + \beta\alpha_2 R_{td_{-2}} + \beta\alpha_3 \Delta R_{max_{-2}} \\ & + \beta^2 R_{s_{-3}}. \end{aligned} \tag{6}$$

The semiannual S and L rate is then the average of equations (5) and (6), and the change in this rate is

$$\frac{R_s + R_{s_{-1}}}{2} - \frac{R_{s_{-2}} - R_{s_{-3}}}{2}$$

Letting the subscript $6t$ represent the current semiannual rate and $3t$ the current quarterly rate, we may represent the change in the 6 month S and L rate, $\Delta R_{s_{6t}}$, as follows:

$$\begin{aligned} \Delta R_{s_{6t}} = {} & \alpha_0(1 + \beta) + \alpha_1[R_{m_{3t}} \\ & + (1 + \beta) R_{m_{3t-1}} + \beta R_{m_{3t-2}}]/2 \\ & + \alpha_2[R_{td_{3t}} + (1 + \beta) R_{td_{3t-1}} \\ & + \beta R_{td_{3t-2}}]/2 + \alpha_3[\Delta R_{max_{3t}} \\ & + (1 + \beta)\Delta R_{max_{3t-1}} \\ & + \beta \Delta R_{max_{3t-2}}]/2 \\ & + (\beta^2 - 1) R_{s_{6t-1}}. \end{aligned} \tag{7}$$

This equation may now be estimated by assuming some β so as to derive the variables in brackets. Since it is unlikely that the β so assumed will equal the β as estimated from the coefficient of the lagged semiannual S and L rate in the regression equation, an iterative procedure was used. First a β equal to 1 was assumed in deriving the bracketed variables. From the resulting regression equation, β was calculated as the square root of 1 plus the coefficient of the lagged S and L rate. This β was then used in calculating the bracketed variables for the next run and so on. After three iterations, the β's converged with a value of .829. The resulting equation is as follows:

$$\begin{aligned} \Delta R_{s_{6t}} = {} & \underset{(2.1573)}{-.6828} + \underset{(5.0979)}{.1495}[\bar{R}_m]/2 \\ & + \underset{(2.6374)}{.0671}[\bar{R}_{td}]/2 \\ & + \underset{(1.5191)}{.0898}[\Delta\bar{R}_{max}]/2 - \underset{(4.1023)}{.3128} R_{s_{6t-1}} \end{aligned} \tag{8}$$

$\bar{R}^2 = .6309 \quad S.E. = .0492 \quad D.W. = 1.5132.$

In equation (8), the numbers in parentheses are t-ratios and the bars over the variables denote weighted averages of the current and lagged values as shown in (7) with β equal to .829. Equation (8) may then be put in terms of equation (4) as follows:

$$\begin{aligned} R_{s_{3t}} = {} & -.3733 + .1495 R_{m_{3t}} \\ & + .0671 R_{td_{3t}} + .0898 \Delta R_{max_{3t}} \\ & + .8290 R_{s_{3t-1}}. \end{aligned} \tag{9}$$

According to this equation, .1710 of the discrepancy between the actual and the desired

[5] R_{td} is available in quarterly form as a result of interpolations performed by Hendershott [4]. R_{td} has been used not only to interpolate the semi-annual S and L rate but also directly in the regressions of Section II.

TABLE 2—INTERPOLATED SERIES ON INTEREST RATES PAID ON SAVINGS AND LOAN SHARES

Quarter	*S* and *L* Rate Unadjusted	*S* and *L* Rate Adjusted
1956–I	3.07	3.06
II	3.07	3.08
III	3.08	3.14
IV	3.16	3.22
1957–I	3.24	3.29
II	3.31	3.34
III	3.39	3.40
IV	3.47	3.46
1958–I	3.54	3.50
II	3.55	3.50
III	3.56	3.47
IV	3.58	3.49
1959–I	3.62	3.52
II	3.65	3.57
III	3.70	3.67
IV	3.78	3.81
1960–I	3.87	3.94
II	3.94	4.02
III	4.01	4.02
IV	4.05	4.03
1961–I	4.09	4.03
II	4.09	4.03
III	4.10	4.05
IV	4.10	4.06
1962–I	4.20	4.18
II	4.21	4.20
III	4.22	4.23
IV	4.23	4.25
1963–I	4.23	4.29
II	4.23	4.28
III	4.23	4.27
IV	4.23	4.27
1964–I	4.23	4.29
II	4.24	4.30
III	4.24	4.30
IV	4.26	4.31
1965–I	4.28	4.32
II	4.29	4.32
III	4.30	4.32
IV	4.34	4.33
1966–I	4.41	4.34
II	4.49	4.42
III	4.60	4.59
IV	4.72	4.71

savings and loan rate is eliminated each quarter, and in the steady state[6]

$$R_s = -2.1830 + .8743R_m + .3924R_{td}.$$

Equation (9) was then used to derive a quarterly *S* and *L* rate series from 1956 to 1966 by first assuming an initial lagged rate such that it and the first calculated rate averaged to the actual first half of 1956 *S* and *L* rate and by then allowing the equation to generate its own lagged values through time. The resulting quarterly series was then adjusted slightly so that every pair of successive quarterly values averaged to the semiannual rates presented in Table 1.[7] Table 2 presents the quarterly *S* and *L* rate both before and after this adjustment.[8] The adjusted series was then used to estimate some of Lee's equations on a quarterly basis as discussed in the following section.

II. *The Demand for Money Using Quarterly Savings and Loan Rates*

Two basic forms of a money demand equation were estimated by Lee. One relates the current quarter's money holdings to permanent income and interest rates and assumes no lags in the adjustment of actual to desired holdings. In multiplicative form,

[6] Although equation (9) was developed primarily to interpolate quarterly a semiannual interest rate series, it should have some economic significance in its own right as a rate setting function on savings and loan shares. In this connection, the fairly slow speed with which R_s moves implies that flows into *S* and *L*'s are likely to be quite volatile as market interest rates change, a conclusion supported by the experience of 1966 and 1967. Also, the steady state properties of R_s imply that a simultaneous increase of 1 percent in both R_{td} and R_m will yield, ultimately, an increase of about 1-1/4 percent in R_s, a not unreasonable result. Although not considered in this note, the effects of the maximum rates which may be paid on savings and loan shares, first established in September, 1966, should be examined for interpolating beyond 1966.

[7] The adjustment process involved first plotting the calculated rate/actual rate ratios for each *semiannual* period and then drawing a free hand polynomial through the plots to develop *quarterly* correction factors such that when these were applied to the original quarterly results, the relevant quarterly values averaged to the actual semiannual values.

[8] The mean absolute difference between the unadjusted and the adjusted series is .0430 and that between the changes of the two series is only .0209.

$$M = a \cdot Y_p^{\alpha} \cdot R^{\beta}, \tag{10}$$

where M is current money holdings, Y_p is permanent income, and R is an interest rate variable.

The other formulation is a multiplicative version of a stock adjustment approach to the demand for money. Thus,

$$M/M_{-1} = (M^*/M_{-1})^{\alpha}, \tag{11}$$

where M, M_{-1}, and M^* are, respectively, this period's money holdings, last period's money holdings, and desired money holdings. It is further specific that

$$M^* = a \cdot Y_p^{\beta} \cdot R^{\gamma}. \tag{12}$$

Lee estimated both formulations by log-linear regressions. The regressions reported below are of this same type, but before discussing these results, some differences between Lee's data and those used here should be mentioned.[9] First, the S and L rate used here is an advertised rate rather than an effective rate as in Lee's work. Secondly, our money stock series is an average of the two months surrounding the end of the quarter rather than of the three months within the quarter. Thirdly, we have used interest rate levels rather than differentials, and finally, our permanent income series is a weighted average of past levels of GNP where the weights are quarterly approximations to the annual ones employed by Friedman [5, pp. 337–38] in computing his permanent income concept.[10] Following Lee, all dollar values are in real per capita terms.[11]

Regressions were run using the S and L rate, the time deposit rate, and both together. Table 3 presents the equations involving no adjustment process, and Table 4 presents the stock adjustment results. Table 3 illustrates clearly the much higher elasticity of the S and L rate than of the time deposit rate as well as the greater significance of the former. Furthermore, when both rates appear, the coefficient of the time deposit rate changes considerably and loses its significance.[12]

The same story is told by Table 4. Moreover, the speed of adjustment is much greater with the S and L rate than with the time deposit rate. Equation (4.1) reveals that in each quarter, the increase of M over M_{-1} eliminates over 60 percent of the percentage difference between M^* and M_{-1}, a rapid rate of adjustment indeed.[13] Nevertheless, the adjustment is not immediate. To compare the speeds of adjustment found in the quarterly equations here with those of Lee's annual equations, it is necessary to take the *average* speed of adjustment of four successive quarters.[14] When this is done, equation

[9] With the exception of the S and L rate, all the data used here are taken from the data bank of the Federal Reserve-MIT econometric model. See Frank de Leeuw and Edward Gramlich [2].

[10] For an example of the use of this variable, see Frank de Leeuw [1, p. 468].

[11] Thus, in the regressions shown, M = real per capita holdings of demand deposits and currency; R_s = the quarterly interest rate paid on savings and loan shares; R_{td} = the quarterly interest rate paid on commercial bank time deposits; Y_P = real per capita permanent GNP with current rather than permanent prices used as deflators.

[12] In an attempt to eliminate the effects of serial correlation, the regressions in Table 3 were transformed by a Cochrane-Orcutt auto-regressive scheme in which it is assumed that the error in period t, u_t, is serially correlated with the error in period $t-1$ and that the remaining error is independently distributed, i.e., $u_t = u_{t-1} + e_t$. This transformation removed the serial correlation in regressions (3.1) and (3.3), but the coefficients of the variables did not change very much. In the case of (3.2) the coefficients changed considerably, but the Durbin-Watson statistic still was too low to be acceptable.

[13] An auto-regressive transformation as discussed in the previous footnote was also applied to the regressions of Table 4. Only in the case of regression (4.2) were the results significantly different, and they are reported here:

$$\begin{aligned} M = {} & \underset{(1.0092)}{.8626} + \underset{(1.3065)}{.1351} Y_p - \underset{(1.9526)}{.0616} R_{td} \\ & + \underset{(8.3200)}{.7189} M_{-1} + .6834 u_{-1} \\ \bar{R}^2 = {} & .9840 \quad S.E. = .0055 \quad D.W. = 1.811 \end{aligned} \tag{4.2a}$$

The evidence of a lagged independent variable in the regressions of Table 4 may call into question the use of this transformation. However, for a similar use of this transformation see de Leeuw and Gramlich [2, p. 31, equation 6].

[14] This, rather than the total adjustment after four quarters, yields the proper measure for comparison with Lee inasmuch as Lee's equations are based on annual averages and not end of year levels.

TABLE 3—REGRESSION EQUATIONS[a] (No lag in adjustment of M)

	Const.	Y_p	R_s	R_{td}	$\bar{R}^2$	S.E.	D.W.
(3.1)	$M=$ 3.4573 (18.2926)	+ .5092 (19.6334)	− .6023 (44.1025)	—	.9862	.0056	.7854
(3.2)	$M=$ 1.1707 (1.5445)	+ .7327 (7.4266)	—	−.2972 (13.3323)	.8750	.0167	.3391
(3.3)	$M=$ 3.0712 (11.7167)	+ .5531 (16.7987)	− .5496 (19.0058)	− .0321 (2.0475)	.9872	.0053	.8537

[a] t-values in parentheses

TABLE 4—REGRESSION EQUATIONS[a] (Lag in adjustment of M)

	Const.	Y_p	R_s	R_{td}	M_{-1}	$\bar{R}^2$	S.E.	D.W.
(4.1)	$M=$ 2.0647 (8.8402)	+.3046 (9.0531)	−.3489 (9.4772)	—	+.3997 (7.1090)	.9938	.0037	1.6049
(4.2)	$M=$ .4664 (1.5190)	+.0437 (.7134)	—	−.0141 (.6627)	+.8796 (14.6988)	.9800	.0067	.9043
(4.3)	$M=$ 2.0578 (8.4376)	+.3080 (6.8714)	−.3484 (9.2866)	−.0014 (.1167)	+.3968 (6.3975)	.9936	.0038	1.5993

[a] t-values in parentheses

(4.1) yields an average annual speed of adjustment of 83.2 percent of the change in M^*, which is close to, but less than the complete adjustment that Lee finds.[15]

Lee's conclusions are thus seen to be confirmed. Savings and loan shares appear to be closer substitutes for money than do time deposits, and Freidman's broader concept of money which includes time deposits but not savings and loan shares may be inappropriate, at least for the postwar period. Also, when the interest rate paid on S and L shares is used in a demand for money equation, discrepancies between desired and actual money holdings are eliminated more rapidly that when the time deposit rate is used.

It should be pointed out however, that in the context of a complete model of the financial sector, the apparent greater speed of adjustment of money holdings using the S and L rate may not be borne out. As equation (9) indicates, the S and L rate itself adjusts rather slowly to other interest rates. Thus, the rapid adjustment of money holdings to the S and L rate is at least partially offset by the sluggish response of that rate to the more direct influences of monetary policy. It is an open question whether equilibrium in financial markets is achieved more quickly with fairly rapid adjustments of asset holdings to slowly moving rates, or slower adjustments of asset holdings to more rapidly moving rates. Simulations of financial models under these two sets of conditions could shed much light on this matter.

[15] This was found by introducing a one-shot 10 percent change in M^* in the equation $M/M_{-1}=(M^*/M_{-1})^{.6003}$ and then solving for M for four successive quarters. The difference between the average M and the base period M was then expressed as a percentage of the change in M^*.

REFERENCES

1. F. DE LEEUW, "A Model of Financial Behavior," in *The Brookings Quarterly Econometric Model of the United States.* J. S. Duesenberry, et. al., eds., Chicago 1965, pp. 465–530.
2. F. DE LEEUW AND E. GRAMLICH, "The

Federal Reserve-MIT Econometric Model," *Fed. Res. Bull.*, Jan. 1968, *54*, 11–40.
3. T. H. LEE, "Alternative Interest Rates and the Demand for Money: the Empirical Evidence," *Am. Econ. Rev.*, Dec. 1967, *57*, 1168–81.
4. P. H. HENDERSHOTT, "Recent Development of the Financial Sector of Econometric Models," *Jour. Finance*, Mar. 1968, *23*, 41–66.
5. M. FRIEDMAN, "The Demand for Money: Some Theoretical and Empirical Results," *Jour. Pol. Econ.* Aug. 1959, *67*, 327–51.

ALTERNATIVE INTEREST RATES AND THE DEMAND FOR MONEY: COMMENT

By MICHAEL J. HAMBURGER*

In his recent study of the influence of various interest rates on the demand for money, T. H. Lee [15] finds that the most important determinants of the aggregate demand for money (defined either to include or to exclude time deposits) are permanent income and the differential between the yield on savings and loan shares and the yield on money. Other variables which were considered but did not contribute to the explanatory power of the model are the lagged stock of money and the differentials between the yields on various types of marketable securities—short-term bonds, long-term bonds, equities—and the yield on money.[1]

On the basis of these results, Lee concludes that " . . . no matter which definition of money is used, the demand for money is highly sensitive to changes in the yield on savings and loan shares, and savings and loan shares are the closest substitutes for money . . . " [15, p. 1175]. In this comment it is shown that Lee's findings depend critically on the use of interest rate differentials. Once this procedure is abandoned and the yield on money is introduced as a separate variable, there is little evidence that savings and loan shares are closer substitutes for money (narrowly defined) than other assets.[2] In addition, the demand for money appears to adjust more slowly to changes in the yield on savings and loan shares than to changes in other rates.

I. *Empirical Results*

Most of the data used in the study were supplied by Lee and are identical to those used in his analysis. The only important modifications are: (1) the introduction of r_{MSB}, the yield on mutual savings bank deposits, and three additional measures of the yield on bonds (r_B)—the rate on U.S. Treasury bills (r_{TB}, a short-term rate), Moody's *Aaa* rate on long-term corporate bonds (r_{Aaa}), and the rate on long-term U. S. Government bonds (r_{US}); (2) the use of both current and lagged values of the independent variables and (3) the use of a separate variable for each interest rate, including the yield on money, instead of interest differentials. Moreover, only the traditional definition of money (demand deposits plus currency) is considered.[3] The symbols used

* The author is special assistant, Federal Reserve Bank of New York. He wishes to thank Anne Roach for her excellent assistance and Tong Hun Lee for supplying most of the data used in the study.

The views expressed are those of the author and do not necessarily reflect those of the Federal Reserve Bank of New York.

[1] When the narrow definition of money is employed, the differential between the yield on time deposits and the yield on money has a modest effect on the demand for money, but the elasticity of this rate is smaller than that of the savings and loan rate.

[2] Similar results are reported in Hamburger [10, 11]. Some of the difficulties associated with using interest rate differentials are discussed in [11].

[3] The reason generally given for including time deposits in the definition of money is that they are much closer substitutes for demand deposits and currency than are other assets. However, as both Lee's results and those shown below indicate, this hypothesis is not supported by the evidence. Questions may also be raised concerning the specification of a number of the variables listed below. Meltzer [17] argues, for example, that it may not be strictly appropriate to deflate variables valued in dollars, i.e., money and income, by population. Such a procedure assumes that the demand function for money is homogenous of the first degree in population alone and disregards the effects of other variables, such as the distribution of population between urban and rural areas. There is also the question of a appropriateness of using total service charges divided by average demand deposit balances as the yield

TABLE 1—LOG-LINEAR REGRESSION EQUATIONS FOR MONEY DEFINED AS DEMAND DEPOSITS PLUS CURRENCY: ANNUAL DATA, 1951–65 (t-statistics in parentheses)

		$\bar{R}^2$ / S	DW
(1)	$M'=3.35+.536Y_P-.634\,(r_{SL}-r_M)$ (6.18) (19.41)	.9881 .0092	1.58
(2)	$M'=3.05+.501Y_P+.156r_M-.374r_{SL}$ (7.76) (4.44) (5.25)	.9935 .0067	2.50
(3.1)	$M'=2.28+.596Y_P+.226r_M-.118r_{MSB}-.022r_{CP}-.088r_E-.208r_{SL-1}$ (6.01) (5.86) (1.74) (2.38) (3.37) (3.15)	.9949 .0060	2.21
(3.2)	$M'=2.24+.597Y_P+.234r_M-.132r_{MSB}-.014r_{TB}-.092r_E-.185r_{SL-1}$ (5.49) (5.62) (1.80) (1.84) (3.22) (2.66)	.9931 .0066	2.15
(3.3)	$M'=3.18+.485Y_P+.216r_M-.093r_{MSB}-.083r_{Aaa}-.091r_E-.171r_{SL-1}$ (4.93) (5.81) (1.39) (2.71) (3.72) (3.00)	.9955 .0056	2.41
(3.4)	$M'=2.97+.505Y_P+.232r_M-.100r_{MSB}-.070r_{US}-.097r_E-.156r_{SL-1}$ (5.00) (6.17) (1.45) (2.45) (3.77) (2.67)	.9951 .0059	2.27
(3.5)	$M'=2.69+.533Y_P+.244r_M-.111r_{MSB}-.049r_{20}-.087r_E-.137r_{SL-1}$ (4.71) (5.78) (1.40) (1.67) (2.96) (2.07)	.9935 .0067	2.65

Note: The subscript (-1) associated with the yield on savings and loan association shares, r_{SL}, indicates that this series is lagged one period.

are as follows:

- M' = Per capita currency and demand deposits held by the public in real terms.
- Y_P = Friedman's permanent per capita net national product in real terms.
- r_M = the yield on money, a weighted average of the yield on demand deposits (the negative of service charges) and currency (zero).
- r_T = the yield on commercial bank time deposits.
- r_{SL} = the yield on savings and loan association shares.
- r_{MSB} = the yield on mutual savings bank deposits.
- r_{CP} = the yield on four- to six-month commercial paper.
- r_{TB} = the yield on three-month Treasury bills.
- r_{20} = the yield on twenty-year corporate bonds.
- r_{Aaa} = Moody's *Aaa* rate on long-term corporate bonds.
- r_{US} = the rate on long-term U. S. Government bonds.
- r_E = Standard and Poor's dividend yield on common stocks. This is the same series as used by Lee, although he refers to it as Moody's.

Table 1 records the results for several log-linear regressions fitted to annual observations of the variables for the period 1951–65. The numbers in parentheses are t-statistics. $\bar{R}^2$ is the squared multiple correlation coefficient adjusted for degrees of freedom, S is the adjusted standard error of the regression, and DW denotes the Durbin-Watson statistic.

Equation (1) reproduces the results obtained by Lee for his equation (1.2).[4] This relationship implies that the logarithm of real per capita money balances is linearly re-

on demand deposits. Brainard [1] argues that this is a very poor proxy for the marginal return to increased holdings of demand deposits.

[4] Since the data used to compute this regression are the same as those employed by Lee, the differences in the results may be attributed to rounding errors.

lated to the logarithm of the difference between the yield on savings and loan shares and the yield on money, and the logarithm of real per capita permanent income. A comparison of these results with those for equation (2), which separates $(r_{SL}-r_M)$ into its two components, clearly demonstrates the disadvantages of using interest rate differentials in the present context. The standard error of equation (1) is approximately 40 percent greater than the standard error of equation (2).

An additional reason for questioning the usefulness of interest rate differentials is provided by equations (3.1) through (3.5), which include several yields and differ among themselves only in terms of the specification of the yield on bonds. They indicate that for the period 1951–65 it is possible to isolate, in addition to r_M, at least three and perhaps four interest rates which have a significant effect on the demand for money, namely, the yields on bonds, equities, savings and loan shares, and mutual savings bank deposits.[5]

The results also bear on the question of the substitution relationships between money and other assets. First, there is little evidence to support the view that short-term interest rates are better indicators of the opportunity cost of holding money than long-term rates.[6] Although the differences in the estimates of $\bar{R}^2$ for equations (3.1) through (3.5) are fairly small, the highest value (.9955) is obtained when the yield on bonds is measured as Moody's *Aaa* rate on long-term corporate bonds.[7] In addition, the results provide some confirmation for my earlier finding [10] that long-term bonds are not much closer substitutes for money than equities. On the margin, equal percentage changes in the yields on these assets will set off similar shifts between money and bonds and between money and equities.

The yield on commercial bank time deposits was also included in the equations. However, as is suggested by the sample results shown below, the addition of this variable introduces a considerable amount of multicolinearity without altering any of the major conclusions of the paper.

$$(3.6)\quad \begin{aligned} M' = 2.50 &+ \underset{(5.45)}{.567Y_P} - \underset{(4.18)}{.277r_M} \\ &- \underset{(1.97)}{.153r_{MSB}} - \underset{(2.48)}{.028r_{CP}} \\ &- \underset{(2.97)}{.116r_B} - \underset{(3.27)}{.224r_{SL-1}} + \underset{(0.95)}{.056r_T} \end{aligned}$$

$$\bar{R}^2 = .9948 \quad S = .0060 \quad DW = 2.08$$

$$(3.7)\quad \begin{aligned} M' = 3.32 &+ \underset{(4.64)}{.473Y_P} - \underset{(5.70)}{.223r_M} \\ &- \underset{(1.38)}{.094r_{MSB}} - \underset{(2.72)}{.095r_{Aaa}} \\ &- \underset{(3.70)}{.094r_B} - \underset{(2.48)}{.225r_{SL-1}} \\ &+ \underset{(0.78)}{.032r_{T-1}} \end{aligned}$$

$$\bar{R}^2 = .9952 \quad S = .0058 \quad DW = 2.70$$

$$(3.8)\quad \begin{aligned} M' = 3.15 &+ \underset{(4.78)}{.492Y_P} - \underset{(6.12)}{.243r_M} \\ &- \underset{(1.45)}{.102r_{MSB}} - \underset{(2.56)}{.086r_{US}} \\ &- \underset{(3.85)}{.102r_B} - \underset{(2.39)}{.224r_{SL-1}} \\ &+ \underset{(0.93)}{.042r_{T-1}} \end{aligned}$$

$$\bar{R}^2 = .9948 \quad S = .0060 \quad DW = 2.57$$

All variables are in natural logarithms and figures in parentheses are *t*-statistics. Similar results are obtained when the yield on bonds is measured as r_{TB} and r_{20} and

[5] The symbol r_{SL-1} is used to reflect the fact that the highest values of $\bar{R}^2$ are obtained when the yield on savings and loan shares is lagged one period. Similar lags were tried for each of the other independent variables, but they did not improve the statistical fit of the equations.

[6] Among those who take this view are Heller [13], Laidler [14], and Teigen [20].

[7] In view of the extremely close correspondence which has been observed between short- and long-term rates in most studies of the term structure of interest rates [12, 18, 19], it is not at all surprising that both rates work almost equally well in explaining the demand for money. The poor performance of the twenty-year bond rate is somewhat disconcerting, however, since it has been used in a number of studies on the demand for money.

Table 2—Regression Coefficients and Other Statistics for Equation (4) Using Various Measures of the Yield on Bonds: Annual Data, 1951–65 (t-statistics in parentheses)

		$\bar{R}^2$ / S	DW
(4.1)	$M' = 2.29 + .595 Y_P + .226 r_M - .118 r_{MSB} - .022 r_{CP} - .088(r_B + r_{SL-1}) - .119 r_{SL-1}$	.9949	
	(6.03) (5.88) (1.74) (2.38) (3.39) (1.63)	.0060	2.21
(4.2)	$M' = 2.24 + .597 Y_P + .235 r_M - .132 r_{MSB} - .014 r_{TB} - .092(r_B + r_{SL-1}) - .093 r_{SL-1}$	.9931	
	(5.50) (5.64) (1.80) (1.84) (3.23) (1.20)	.0065	2.15
(4.3)	$M' = 3.18 + .484 Y_P + .217 r_M - .093 r_{MSB} - .083 r_{Aaa} - .091(r_B + r_{SL-1}) - .079 r_{SL-1}$	.9955	
	(4.94) (5.83) (1.39) (2.72) (3.73) (1.25)	.0056	2.40
(4.4)	$M' = 2.97 + .505 Y_P + .232 r_M - .100 r_{MSB} - .070 r_{US} - .097(r_B + r_{SL-1}) - .059 r_{SL-1}$	.9951	
	(5.02) (6.19) (1.45) (2.45) (3.79) (0.91)	.0059	2.26
(4.5)	$M' = 2.69 + .533 Y_P + .244 r_M - .111 r_{MSB} - .049 r_{20} - .088(r_B + r_{SL-1}) - .050 r_{SL-1}$	.9935	
	(4.72) (5.80) (1.39) (1.67) (2.97) (0.66)	.0067	2.65

Note: The subscript (−1) associated with the yield on savings and loan association shares, r_{SL}, indicates that this series is lagged one period.

when both a short- and long-term rate are included in the same equation.

We turn next to the hypothesis that savings and loan shares are the closest substitute for money. If this were the case, we would expect the absolute value of the elasticity of the demand for money with respect to the yield on these claims to be greater than the elasticitiy associated with the yield on any other asset, particularly the yield on equities. This implies that the absolute values of the coefficients of r_{SL-1} in equations (3.1) through (3.5) should be significantly greater than the absolute values of the coefficients of r_B. It can be shown that this will be true, if and only if the estimates of b_6 in equation (4) are negative and are significantly different from zero.

$$
\begin{aligned}
(4) \qquad \log M' = a &+ b_1 \log Y_p + b_2 \log r_M \\
&+ b_3 \log r_{MSB} + b_4 \log r_B \\
&+ b_5(\log r_B + \log r_{SL-1}) \\
&+ b_6 \log r_{SL} + \epsilon
\end{aligned}
$$

where r_B denotes the yield on bonds and ϵ is a random error term.[8]

[8] This test is critically dependent upon the specification of the variables. For example, if the variance of the yield on savings and loan shares is overstated (which seems unlikely), the estimate of b_6 will be biased towards zero.

Table 2 presents the estimates of equation (4) for various measures of r_B. As might be expected, the results are almost identical to those for equations (3.1) through (3.5). However, they demonstrate that, once r_M is included in the money demand function as a separate independent variable, there is little or no evidence that savings and loan shares are closer substitutes for money than equities. The estimates of b_6 always have the anticipated signs (negative) but are never statistically significant at the 5 percent level.

Finally, there is the question of the speed with which aggregate money balances adjust to changes in interest rates. Lee finds that the coefficient of the lagged dependent variable in the demand function for money becomes statistically insignificant whenever the yield on savings and loan shares and/or the yield on time deposits are introduced into the equation [15, p. 1177]. From this he concludes that the lag in the adjustment of actual money holdings in response to changes in the yields on savings accounts is quite short or perhaps nonexistent. The tests conducted in this study (results not shown) confirm the findings concerning the importance of the lagged dependent variable in the demand function for money (fitted to annual data). However, as the

subscript (−1) associated with r_{SL} in Tables 1 and 2 shows, the demand for money may, in fact, adjust more slowly to changes in the yield on savings and loan shares than to changes in other rates.[9]

II. *Conclusions*

The results presented above support my earlier conclusions [10, 11] concerning the relative rates of substitution between money (narrowly defined) and other assets, and hence are not consistent with the views of Gurley and Shaw [8, 9], Lee [15, 16], Duesenberry [6], Heller [13], and Laidler [14]. Generally speaking, the results fail to reveal any particular class of assets that may be viewed as either very close or very poor substitutes for money.[10] There is little evidence to suggest that short-term bonds (U. S. Treasury bills) are closer substitutes for money than long-term bonds. Similarly, there is little evidential support for the hypothesis that equities (i.e., claims against the earnings and physical assets of business firms) are much poorer substitutes for money than either bonds or the liabilities of financial intermediaries. Hence, we cannot agree with Lee that monetary economists ought to pay particular attention to the developments of nonbank financial intermediaries and to their role in affecting the demand for money.

REFERENCES

1. W. C. BRAINARD, A review of *The Demand for Liquid Assets: A Temporal Cross-Section Analysis* by E. L. Feige. *Am. Econ. Rev.*, March 1965, *55*, 241–43.
2. K. BRUNNER AND A. H. MELTZER, "Predicting Velocity: Implications For Theory and Policy," *Jour. Fin.*, May 1963, *18*, 319–54.
3. B. C. COHEN AND G. G. KAUFMAN, "Factors Determining Bank Deposit Growth by State: An Empirical Analysis," *Jour. Fin.*, March 1965, *20*, 59–70.
4. F. DE LEEUW, "A Model of Financial Behavior," in J. Duesenberry, G. Fromm, L. Klein and E. Kuh, eds., *The Brookings Quarterly Econometric Model of the United States*, Chicago 1965.
5. ———, "The Demand for Money—Speed of Adjustment, Interest Rates, and Wealth," in G. Horwich, ed., *Monetary Process and Policy: A Symposium*, Homewood 1967.
6. J. DUESENBERRY, "The Portfolio Approach to the Demand for Money and Other Assets," *Rev. Econ. Stat.*, Suppl. Feb. 1963, *45*, 9–31.
7. I. FRIEND, "The Effects of Monetary Policies on Nonmonetary Financial Institutions and Capital Markets," in Comm. on Money and Credit, *Private Capital Markets*, Englewood Cliffs, 1964.
8. J. G. GURLEY AND E. S. SHAW, "Financial Intermediaries and the Saving-Investment Process," *Jour. Fin.*, March 1956, *11*, 257–76.
9. ———, *Money in a Theory of Finance*. Washington, D.C., 1960.
10. M. J. HAMBURGER, "The Demand for Money by Households, Money Substitutes, and Monetary Policy," *Jour. Pol. Econ.*, Dec. 1966, *74*, 600–23.
11. ———, "The Household Demand for Financial Assets," *Econometrica*, Jan. 1968, *36*, 97–118.
12. ——— AND C. M. LATTA, "The Term Structure of Interest Rates: Some Additional Evidence," *Jour. Money Credit Banking*, March 1969.
13. H. R. HELLER, "The Demand for Money: The Evidence from the Short-Run Data," *Quart. Jour. Econ.*, May 1965, *79*, 291–303.
14. D. LAIDLER, "The Rate of Interest and the Demand for Money," *Jour. Pol. Econ.*, Dec. 1966, *74*, 543–55.
15. T. H. LEE, "Alternative Interest Rates and the Demand for Money: The Empirical Evidence," *Am. Econ. Rev.*, Dec. 1967, *57*, 1168–81.
16. ———, "Substitutability of Nonbank Intermediary Liabilities for Money: The Empirical Evidence," *Jour. Fin.*, Sept. 1966, *21*, 441–57.
17. A. H. MELTZER, "The Demand for

[9] For a more general discussion of the use of lagged independent variables, see Hamburger [10, 11].

[10] Other writers who have obtained similar results include Brunner and Meltzer [2], Cohen and Kaufman [3], de Leeuw [4, 5], Friend [7], and Meltzer [17].

Money: The Evidence from the Time Series," *Jour. Pol. Econ.*, June 1963, *71*, 219–46.
18. F. Modigliani and R. Sutch, "Debt Management and the Term Structure of Interest Rates: An Empirical Analysis of Recent Experience," *Jour. Pol. Econ.*, Suppl. Aug. 1967, *75*, 569–89.
19. A. Okun, "Monetary Policy, Debt Management and Interest Rates: A Quantitative Appraisal," in Comm. on Money and Credit, *Stabilization Policies*, Englewood Cliffs, 1963.
20. R. L. Teigen, "Demand and Supply Functions for Money in the United States: Some Structural Estimates," *Econometrica*, Oct. 1964, *32*, 476–509.

ALTERNATIVE INTEREST RATES AND THE DEMAND FOR MONEY: REPLY

By Tong Hun Lee*

While Harvey Galper, using quarterly data, confirms most of the findings in my earlier paper [7], Michael Hamburger attempts to show that his alternative method using my annual time-series data does not confirm my findings. A further analysis of the annual data, however, reveals that Hamburger is not correct. Moreover, a similar analysis with Galper's quarterly data shows results consistent with my original findings, thus again rejecting Hamburger's contention. On the other hand, although I agree with most of Galper's results, I will show that Galper's quarterly data give results more closely related to my original results than Galper's analysis indicates. Therefore, it can be concluded that my previous findings are reaffirmed either on the basis of annual time-series data or on the basis of quarterly data.

* The author is professor of economics at the University of Wisconsin-Milwaukee, and a member of the Social Systems Research Institute at the University of Wisconsin, Madison. He is grateful to Frank de Leeuw for kindly providing the data on the time deposit rate, which is used in this paper, following Galper. He is also indebted to Leon Schur for his helpful comments and to Paul Tsao for his programming assistance. Thanks are due to the Graduate School of the University of Wisconsin-Milwaukee for its financial support of this research.

I

Contrary to one of the conclusions in my paper, Hamburger argues that the liabilities of nonbank financial intermediaries are not closer substitutes for money than are other financial assets and equities in particular. To support his view, he presents the results obtained by regressing money stock on a lagged savings and loan rate among other variables (in real per capita and logarithmic terms). Hamburger has supposedly chosen the lagged S & L rate instead of the current S & L rate, since the use of the lagged S & L rate gave a better fit. Although the differences in $\bar{R}^2$s are extremely small and only half of his relevant regressions show better fits, to be discussed later, he nevertheless chooses a set of regression equations with the lagged S & L rate rather than that with the current S & L rate. According to Hamburger, the reason for using a lagged interest rate is to capture the time lag involved in the demand for money in response to changes in the corresponding interest rate. Since I employed the stock adjustment model in my previous paper, possible time lags are already taken into account in such a formulation if time lags in response to various interest rates are more or less the same. Hamburger apparently feels that time lags in response to changes in various interest rates are different. Although I am sympathetic to this view, I do not think that annual time-series data as opposed to quarterly data provide a basis for estimating such a detailed structure of time lags, especially when differences in $\bar{R}^2$s are negligibly small. This is one of the reasons why I employed the stock adjustment model and also cautioned about the use of annual time-series data in estimating time lags [7, p. 1178, fn. 11]. However, I will not press this argument and will go on to analyze the data along the lines suggested by Hamburger.

Hamburger introduces the yield on mutual savings bank deposits in his regression analysis of the money demand function, but

none of the relevant regression coefficients in his equations shows statistical significance at the conventional probability level. Thus, one can perform Hamburger's analysis by omitting the yield on mutual savings bank deposits. This makes sense, particularly because the S & L rate included in the analysis reflects the influence of more important nonbank intermediary liabilities on the demand for money. Consequently, as the first stage of stepwise regression analysis, I define the demand for money as a function of the S & L rate, the yield on equities, and any one of several short-term and long-term rates, while introducing permanent income and the yield on money into the demand function.[1]

In Table 1, using Hamburger's symbols, the regressions of the money demand function with the current S & L rate are reported in equations (1)–(5), whereas the regressions with one year lagged S & L rate are given in equations (6)–(10).[2] Since $\bar{R}^2$ in each of the former equations is higher than that in the corresponding latter ones, one has to choose in accordance with Hamburger's procedure the regressions with the current S & L rate. Thus, the results contradict Hamburger's assertion that the demand for money adjusts more slowly to changes in the yield on savings and loan shares than to changes in other rates. Moreover, in every regression with the current S & L rate through equations (1) and (5), the elasticity of the S & L rate is significantly greater than the elasticity of any other rate, including the yield on equities. The results, therefore, confirm my original findings. Even in the regression equation (8), which Hamburger would prefer and which is the best among the equations with the lagged S & L rate, the elasticity of the S & L rate is greater than that of the yield on equities at about the .09 significance level. Thus, Hamburger is clearly incorrect in arguing that nonbank intermediary liabilities are not closer substitutes for money than are equities. A further searching for the best fitted regression with the above set of independent variables provides equation (11) reported in Table 1. The result again confirms my original conclusion, thereby rejecting Hamburger's contention.

In the second stage of my stepwise regression analysis, the yield on mutual savings bank deposits is added to other independent variables already included in the demand function. Although not reported in detail due to space limitation, the results do not clearly indicate that Hamburger's regressions in his Table 1 with the lagged S & L rate are superior to the equivalent regression equations with the current S & L rate. His regression equations in (3.1) and (3.2) of his Table 1 are inferior to the corresponding equations with the current S & L rate while his regression equations in (3.3) and (3.4) of his Table 1 are somewhat better than the corresponding equations with the current S & L rate.[3] His last regression equa-

[1] Hamburger argues that a separate introduction of the yield on money into the demand function of money has resulted in his findings, which differ from my original findings. That this is not true can readily be seen from the present analysis, which follows Hamburger's procedure. In fact, I do not have any quarrel on the use of the yield on money as a separate regressor, since I had already used it in my earlier analysis with various cross-sectional data [8]. The reason that I did not use this type of formulation with annual time-series data was to lessen the problem of intercorrelations among interest rate variables. Also to alleviate such a problem, I used in my previous analysis and will again use in the present analysis the stepwise regression method, whereas Hamburger does not appear to be concerned about it since he uses a given set of interest rate variables that suits best his conclusions.

[2] In Table 1, the symbol r_M is the yield on money but it is defined as Hamburger's r_M with the sign reversed. The reason for this difference in definitions is to avoid the negative sign in Hamburger's r_M in the regression analysis. Hamburger actually used the present definition of r_M in his regression analysis, but he reversed the sign of the regression coefficient on r_M after having obtained his regression. Since this transformation is not strictly correct algebraically, the present results in Table 1 are reported without such a transformation.

[3] I found some minor numerical errors in Hamburger's computations, although they do not affect his analysis in any significant way. The permanent income figures he used in his calculations are greater in size by .01 than the figures that I had provided to him. The S & L rates which he obtained elsewhere have two significant digits by rounding off the third digit for the 1950–57 figures, while they show three significant digits for the rest of the period of analysis. Also he takes 4.21 as the 1965 S & L rate while it should have been 4.22. Since the resulting computational errors are small, I will not reproduce these results. The $\bar{R}^2$s that are corrected for Ham-

TABLE 1—ANNUAL LOG-LINEAR REGRESSION EQUATIONS FOR MONEY DEFINED AS DEMAND DEPOSITS PLUS CURRENCY: ANNUAL DATA, 1951–65 (t-statistics in parentheses)

	$\bar{R}^2$	DW	t-ratio*
(1) $M'=3.34+.459Y_P-.204r_M-.009r_{CP}-.330r_{SL}-.056r_E$ (6.61) (5.12) (1.18) (4.78) (2.14)	.9955	2.90	3.29
(2) $M'=3.32+.462Y_P-.206r_M-.006r_{TB}-.331r_{SL}-.057r_E$ (6.54) (5.08) (1.05) (4.73) (2.15)	.9954	2.83	3.24
(3) $M'=3.71+.412Y_P-.205r_M-.040r_{Aaa}-.292r_{SL}-.062r_E$ (5.45) (5.23) (1.32) (3.86) (2.39)	.9957	2.89	2.55
(4) $M'=3.62+.423Y_P-.212r_M-.036r_{US}-.289r_{SL}-.065r_E$ (5.78) (5.36) (1.30) (3.77) (2.45)	.9957	2.86	2.43
(5) $M'=3.39+.451Y_P-.209r_M-.004r_{20}-.328r_{SL}-.058r_E$ (5.77) (4.64) (.12) (3.33) (2.03)	.9949	2.53	2.39
(6) $M'=3.03+.485Y_P-.256r_M-.023r_{CP}-.260r_{SL-1}-.087r_E$ (4.98) (6.13) (2.27) (3.36) (2.81)	.9930	2.03	1.95
(7) $M'=3.08+.473Y_P-.270r_M-.015r_{TB}-.242r_{SL-1}-.091r_E$ (4.53) (6.22) (1.80) (2.96) (2.76)	.9918	1.97	1.61
(8) $M'=3.86+.392Y_P-.225r_M-.102r_{Aaa}-.219r_{SL-1}-.089r_E$ (5.32) (6.24) (3.62) (3.75) (3.65)	.9955	2.73	1.93
(9) $M'=3.67+.408Y_P-.247r_M-.087r_{US}-.204r_{SL-1}-.097r_E$ (5.21) (6.93) (3.24) (3.29) (3.71)	.9949	2.40	1.51
(10) $M'=3.37+.436Y_P-.262r_M-.068r_{20}-.185r_{SL-1}-.084r_E$ (4.90) (6.74) (2.51) (2.63) (2.83)	.9934	2.53	1.26
(11) $M'=3.18+.478Y_P-.180r_M-.014r_{TB-1}-.328r_{SL}-.026r_E$ (7.64) (4.77) (2.05) (5.36) (.92)	.9965	2.49	4.04
(12) $M'=3.31+.457Y_P-.156r_M+.082r_{MSB}-.019r_{TB-1}-.395r_{SL}+.024r_{E-1}$ (5.49) (5.33) (1.12) (3.27) (5.21) (1.06)	.9965	2.21	5.16
(13) $M'=2.80+.530Y_P-.158r_M+.078r_{MSB}-.033r_{CP-1}+.096r_{Aaa-1}-.517r_{SL}+.010r_{E-1}$ (5.03) (4.96) (.93) (2.01) (1.62) (5.07) (.383)	.9960	2.87	5.19

* t-ratio to determine significance of difference between partial regression coefficients: M' on S and L rate; M' on equity yield.

tion in (3.5) of his Table 1 gives the same size of $\bar{R}^2$ as the corresponding equation with the current S & L rate. The only reason that Hamburger could have had for using his regressions with the lagged S & L rate is that his regression (3.3) gives the highest correlation among the above regression equations. By following his rule, however, the highest $\bar{R}^2$ can be obtained not by lagging the S & L rate, but by lagging other rates while retaining the current S & L rate. The regression that gives the highest correlation is reported in equation (12) of Table 1. This result, while rejecting Hamburger's use of

burger's regressions in his Table 1 are as follows: .9949 for his equation (3.1), .9941 for (3.2), .9962 for (3.3), .9956 for (3.4) and .9944 for (3.5). On the other hand, the $\bar{R}^2$s that are computed by using the current S & L rate instead of the lagged S & L rate in Hamburger's equations are .9951, .9951, .9953, .9953 and .9944 respectively for the corresponding equations listed above. In the comparison of $\bar{R}^2$ referred to in the text, I am using the $\bar{R}^2$ computed after correcting the numerical errors in Hamburger's data. In all other equations with annual data, I am also using the corrected data.

the lagged S & L rate, confirms my previous conclusion that the S & L rate is the most important interest rate variable affecting the demand for money.

Finally, I used the S & L rate, the yield on equities, the yield on mutual savings bank deposits, a short-term rate, and a long-term rate as independent variables, in addition to permanent income and the yield on money. Since Hamburger claims that a particular long-term rate may give a better fit than other long-term rates, I used each of his long-term rates in turn in the final regression analysis. The best equation, as selected by use of Hamburger's criterion, is given in equation (13) of Table 1. The result again indicates the same conclusion that the non-bank intermediary liabilities are the closest substitutes for money.

This conclusion is further supported by the results from quarterly data to be discussed in the next section. Hamburger, therefore, is incorrect in his comment and makes erroneous conclusions even on the basis of his own ground rules.

II

Galper supports most of the findings in my earlier paper [7] by employing the quarterly interest rates on savings and loan shares that he generates by interpolating the semiannual advertised rates. The only difference, according to Galper, is that his quarterly equation gives a small but significant coefficient of the lagged money stock, whereas my annual equation showed a statistically insignificant coefficient. A further examination of his data, however, reveals that this difference too disappears.

For convenience of obtaining data, Galper uses different definitions of variables for money, permanent income, and interest rates. For a strict comparison with my original paper, however, one should define money stock (currency plus demand deposits) as an average of the three months within the quarter, and permanent income as the series obtained by using Friedman's quarterly weights.[4] Since there are no quarterly data on the effective interest rate on savings and loan shares, one may utilize Galper's interpolated series, but with a caution to be noted later. The quarterly regressions of money demand function in these variables (in real per capita and logarithmic terms) with and without the lagged money stock, M_{-1}, are given in (1) and (4) of Table 2 below. While these regression equations fit better than the corresponding Galper's equations, the estimate, .319, of the coefficient of M_{-1} is much smaller than Galper's estimate of .397.

The above estimate of time lag, as small as it is, contains an upward bias. The reason for the upward bias in the estimated coefficient of M_{-1} stems from the fact that the estimating equation uses an advertised rate rather than an effective rate on the savings and loan shares. Since an advertised rate is usually associated with such other terms as holding period of savings balances and frequency of compounding, the advertised rate alone does not reflect the return on holdings of saving and loan shares for the relevant period. Such a difference in the specification of the rate of return on savings and loan shares would have been caught up by an estimated coefficient of M_{-1}. However, if an advertised rate alone had to be used in absence of information on other terms, its lagged variable rather than the current variable may be more appropriate in estimating the quarterly money demand function. There are two reasons for this. First, an announced change in advertised rate may be felt among depositors after some delay, due to the time involved in informing the public not only of the rate change but also of other terms. Second, the quarterly data give a better basis for obtaining time lags of this type than do the annual data. In other words, Hamburger's method of analysis should have been applied to the quarterly data rather than to the annual data. For these reasons, in estimating the quarterly money demand function, I used the S & L rate and the time deposit rate with different lags while omitting M_{-1}. The resulting equation that gives the maximum $\bar{R}^2$ is reported in equation (2) of Table 2. Note that this equation with the S & L rate

[4] The quarterly weights can be derived by Friedman's method [1, pp. 142–152] but with the weighting parameter, β, being equal to .1.

TABLE 2—QUARTERLY LOG-LINEAR REGRESSION EQUATIONS FOR MONEY DEFINED AS DEMAND DEPOSITS PLUS CURRENCY: QUARTERLY DATA, 1956–66 (t-statistics in parentheses)

	$\bar{R}^2$	DW	t-ratio*
(1) $M'=2.73+.583Y_P-.064r_T-.430r_{SL}$	.9901	.763	
(13.60) (3.31) (11.48)			
(2) $M'=3.48+.487Y_P-.030r_T-.459r_{SL-1}$	.9920	1.12	
(12.02) (1.62) (14.51)			
(3) $M'=4.16+.412Y_P-.001r_{T-1}-.475r_{SL-1}+.007r_{CP-1}-.054r_{Aaa-1}-.028r_{E-1}$	.9954	1.23	17.92
(11.34) (.04) (20.26) (1.38) (3.49) (4.27)			
(4) $M'=2.37+.336Y_P-.010r_T-.345r_{SL}+.319M'_{-1}$	.9940	1.13	
(5.20) (.49) (9.60) (4.53)			
(5) $M'=4.13+.413Y_P-.018r_T-.420r_{SL-1}+.109M'_{-1}$	.9922	.884	
(5.94) (.84) (9.64) (1.29)			
(6) $M'=3.67+.369Y_P+.008r_{T-1}-.429r_{SL-1}+.001r_{CP-1}-.049r_{Aaa-1}-.028r_{E-1}+.125M'_{-1}$	.9957	1.30	17.08
(7.56) (.53) (11.94) (.20) (3.18) (4.37) (1.68)			

Note: The symbols are defined as those in Hamburger's paper with the exception that r_T and r_{T-1} are the interest rates on commercial bank time deposits of current quarter and one-quarter lag respectively.

* t-ratio to determine significance of difference between partial regression coefficients: M' on S and L rate; M' on equity yield.

lagged one-quarter exhibits a better fit than equation (1) in Table 2. More interestingly, computing the same equation, but with M_{-1} in regression (5), now shows an insignificant coefficient of M_{-1}.

The above results, therefore, indicate that the speed of adjustment estimated by Galper tends to overstate the time lag involved in the demand for money. While Galper's estimate of .397 for the coefficient of M_{-1} implies about 84 percent adjustment in the first two quarters, the present estimate indicates the complete adjustment within two quarters in response to changes in interest rates. Thus, the time lag implied by the quarterly series is shorter than Galper's analysis indicates and is quite small, as I implied in my earlier paper [for example, see [7, p. 1178, fn. 11]). In addition, the estimated coefficient of M_{-1} from quarterly data just as that from annual data does not turn out to be significant. As I indicated in my previous paper, this result has an important implication for rejecting Tucker's explanation of a speedy income adjustment to money-supply changes [9].[5]

When looking at the results in Section II, one might be tempted to observe that, though not supported with annual data, Hamburger may still be correct in arguing the importance of a lagged S & L rate on the demand for money. Contrary to Hamburger's assertion, however, the S & L rate refers to an advertised rate showing only a one-quarter lag, and other interest rates are similarly lagged. Moreover, the regression equation (3) of Table 2, in which a set of interest rate variables similar to Hamburger's interest rate variables is used, indicates that savings and loan shares are closer substitutes for money than are other assets. Although a long-term rate turns out to be significant, the elasticity of the S & L rate is about nine times greater than the

[5] Galper correctly points out that despite all the implications of the estimated money demand function for time lags, the greater speed of adjustment of money holdings using the savings and loan rate may not be borne out in a complete model of financial behavior. However, the discussions in my original paper, excepting those related to Tucker's explanation, are limited to the question of how rapidly the public responds to changes in permanent income and interest rates rather than to changes in the basic monetary policy variables. A knowledge of the former, of course, precedes any understanding of the latter.

elasticity of the long-term rate, which is only .054 in size. The null hypothesis that the difference between the two elasticities is zero is statistically rejected with the t-statistic of 13.01. The elasticity of the S & L rate is also greater than that of the yield on equities with the t-statistic of 17.92 for the difference between the two elasticities.

At this point, I must repeat the basic objective of my previous paper, because Hamburger's comments have diverted attention from it. Contrary to Hamburger's implication in his comment, I have not advocated that interest rates other than the S & L rate do not affect the demand for money. While admitting the possible influence of a whole spectrum of interest rates on the demand for money, the principal question of the paper was: "Which interest rate approximates 'the' rate of interest and exerts the most significant influence on the demand for money?" [7, p. 1168]. The answer was, of course, given in a similar context with the various implications of this finding. The reason for not attempting to find the exact set of interest rate variables affecting the demand for money is that it is difficult to determine such a set at least from annual time-series data due to high correlations among interest rate variables. Although Hamburger claims that his method yields an identification of the separate effects of several other interest rate variables on the demand for money, his claim does not hold true with annual time-series data. To see this, note that in the final regression with annual data in equation (13) of Table 1, other interest rates, excluding the S & L rate and the yield on money, are not statistically significant, while some of the other rates turn out to be significant in certain regression equations at an earlier stage of stepwise regression analysis. My previous analysis yielded similar results. Thus, the quantitative importance of other rates in affecting the demand for money is not so clear from the analyses of annual data. Note, however, that the S & L rate is highly significant in any equation of the stepwise regression analysis. What is clear from the annual data, therefore, is that the S & L rate is the most important interest rate variable affecting the demand for money, a result which had not been found in prior studies.[6] Although a further analysis of the quarterly data provides some insight into the importance of other rates, the S & L rate turns out again to be the most significant interest rate variable affecting the demand for money.

The results from both quarterly and annual data, therefore, clearly support my original findings that savings and loan shares are more important than other assets as money substitutes. Thus, the results substantiate the substitution hypothesis of Gurley and Shaw [2] [3] [4] [5], indicating that the nonbank intermediary liabilities are close substitutes for money. Finally, since Galper confirms my original findings and my further analysis does so to an even greater degree, one may conclude that the lags of the demand for money in response to interest rate changes are small when the yield on savings and loan shares is incorporated into the demand function for money.

REFERENCES

1. M. Friedman, *A Theory of the Consumption Function*. Princeton 1957.
2. J. G. Gurley and E. S. Shaw, "Financial Aspects of Economic Development," *Amer. Econ. Rev.*, Sept. 1955, *45*, 515–38.
3. ———, "Financial Intermediaries and the Saving-Investment Process," *Journ. Fin.*, May 1956, *11*, 257–76.
4. ———, *Money in a Theory of Finance*, Washington, D.C., 1960.
5. ———, "The Growth of Debt and Money in the United States, 1800–1950: A Suggested Interpretation," *Rev. Econ. Stat.*, Aug. 1957, *37*, 250–62.
6. M. J. Hamburger, "The Demand for Money by Households, Money Substitutes, and Monetary Policy," *Journ. Pol. Econ.*, Dec. 1966, *74*, 600–23.
7. T. H. Lee, "Alternative Interest Rates and the Demand for Money: The Empiri-

[6] When rejecting the above conclusion, Hamburger also cites his finding in an earlier paper [6]. Although this paper cannot be examined in detail here, it is worthwhile pointing out that he did not use the S & L rate in his analysis and hence, his analysis does not have much to say about the influence of the S & L rate on the demand for money.

cal Evidence," *Amer. Econ. Rev.*, Dec. 1967, *57*, 1168–81.
8. ———, "Substitutability of Non-Bank Intermediary Liabilities for Money: The Empirical Evidence," *Journ. Fin.*, Sept. 1966, *21*, 441–57.
9. D. P. TUCKER, "Dynamic Income Adjustment to Money-Supply Changes," *Amer. Econ. Rev.*, June 1966, *56*, 433–49.

TARIFFS AND TRADE IN GENERAL EQUILIBRIUM: COMMENT

By RONALD W. JONES*

A standard result in the theory of tariffs has been challenged by B. Södersten and Karl Vind in a recent issue of this *Review* [5]. After an explicit development of the general equilibrium model appropriate or this analysis, they conclude that the classic result on the possible effect of tariffs on domestic prices achieved in 1949 by Lloyd Metzler, [3] and [4], is incorrect. In the opinion of Södersten and Vind, a tariff levied on imports can never result in a lowering of the relative domestic price of the imported commodity.

This is a surprising conclusion, the more so as Metzler's result has gone unchallenged for almost twenty years. I propose to demonstrate that Södersten and Vind are themselves the guilty parties in this issue and that Metzler's result is indeed correct. In my opinion the general equilibrium analysis used by Södersten and Vind, essentially a model with eleven equations and eleven unknowns, is unnecessarily cumbersome for the purpose. Therefore in the first section of this paper I shall offer a simpler treatment of the same model to prove the important theorems in the standard theory of tariffs. In the concluding section, I shall point out where they have allowed errors to lead them to incorrect results. The reasons cited by Södersten and Vind for their departure from the standard Metzler result are irrelevant to the argument, for the real culprit lurks instead in a footnoted assumption about the behavior of consumption with a change in relative prices.

* The author is professor of economics at the University of Rochester.

I. *The Analysis of Tariffs*

The setting for the analysis of tariffs is the same as that used by Södersten and Vind, although I shall employ a different terminology. Two countries, 1 and 2, are involved in trading two commodities, A and B, produced under conditions of increasing opportunity costs. Production is carried on under conditions of perfect competition, subject to tariffs levied on imports by both countries. Let (A_i, B_i) represent the consumption bundle, and (a_i, b_i) the production bundle, in country i. Country 1 is assumed to import commodity B, with M_1 equal to $B_1 - b_1$. Similarly country 2's imports are denoted by $M_2 \equiv A_2 - a_2$. Prices differ between countries because of tariffs. (p_a, p_b) represents domestic prices in country 1 while (π_a, π_b) represents prices in country 2. The ad valorem tariff of country i is t_i, and let T_i represent one plus the tariff rate. Thus p_b equals $T_1\pi_b$ and π_a is given by $T_2 p_a$. As usual, in these models only relative prices count. Let commodity A serve as *numeraire* so that B's relative price ratio in country 1 is $p \equiv p_b/p_a$ and in country 2 is $\pi \equiv \pi_b/\pi_a$. Finally, consider the world terms of trade, the untaxed price of B in A units, β. Thus p equals $T_1\beta$ and π equals β/T_2.

Tariff analysis is often portrayed by means of a diagram showing how a tariff serves to shift a country's offer curve, resulting in a new intersection with the other country's offer curve and a new equilibrium world terms of trade. In this spirit consider the representation shown by (1) and (2),

$$M_1 = M_1(\beta; T_1) \tag{1}$$

$$M_2 = M_2\left(\frac{1}{\beta}; T_2\right) \tag{2}$$

where each country's demand for imports is shown as a function of the *untaxed* relative price of those imports and of that country's tariff rate. Differentiate these expressions totally to obtain (3) and (4), where a caret

"^" over a variable denotes the relative change in that variable (e.g., $\hat{M}_1 \equiv dM_1/M_1$):

$$(3)\qquad \hat{M}_1 = -\epsilon_1\hat{\beta} - \bar{\epsilon}_1\hat{T}_1$$

$$(4)\qquad \hat{M}_2 = \epsilon_2\hat{\beta} - \bar{\epsilon}_2\hat{T}_2$$

where

$$\epsilon_1 \equiv -\frac{\beta}{M_1}\frac{\partial M_1}{\partial \beta}; \qquad \epsilon_2 \equiv -\frac{\left(\frac{1}{\beta}\right)}{M_2}\frac{\partial M_2}{\partial\left(\frac{1}{\beta}\right)}$$

and

$$\bar{\epsilon}_i \equiv -\frac{T_i}{M_i}\frac{\partial M_i}{\partial T_i}$$

The ϵ_i are the traditional elasticities of demand for imports along the offer curves. These elasticities are defined so as to be positive. The $-\bar{\epsilon}_i$ define the change in each country's import demand as of constant world terms of trade. The reason I choose to label these $-\bar{\epsilon}_i$ will become apparent later.

Much of the complexity surrounding a general equilibrium analysis of tariffs arises from the necessity of breaking down the ϵ_i and $\bar{\epsilon}_i$ into more fundamental components. At a constant rate of tariff, a change in the world terms of trade induces a change in the consumption of importables. For there must also be a change in the domestic terms of trade and therefore substitution and income effects. In addition, production will generally alter along the production-possibility schedule. In terms of the fundamental determinants of a country's level of imports, the crucial variables are the domestic price ratio and real income.

Consider the analysis for country 1, whose demand for imports is shown functionally in equation (5).

$$(5)\qquad M_1 = B_1(p, y_1) - b_1\left(a_1\left(\frac{1}{p}\right)\right)$$

The term "y_1" denotes country 1's "real income" measured in units of A, the *numeraire*. It is only the change in real income that is of concern, and this can be defined as the domestic price-weighted sum of the changes in consumption as in (6).[1]

$$(6)\qquad dy_1 \equiv dA_1 + pdB_1$$

The budget constraint for country 1, shown in (7), states that at domestic prices the value of consumption is limited by the value of production and the tariff

$$(7)\quad A_1 + pB_1 = a_1 + pb_1 + (T_1 - 1)\beta M_1$$

revenue.[2] By differentiating this totally, and making use of the definition of the real income change shown in (6), it is possible (in (8)) to show the sources of any change in real income.

$$(8)\quad dy_1 = -M_1dp + d\{(T_1 - 1)\beta M_1\}$$

In deriving (8) use has been made of the general production relationship $\{da_1 + pdb_1\} = 0$. If country 1 is incompletely specialized this reflects the tangency of the domestic price line and the transformation schedule. If country 1 is completely specialized to its export commodity and remains at the corner of its transformation schedule, da_i and db_i are each zero.

Although the expression for dy_1 in (8) is not yet in the form most useful for tariff analysis, it does reveal an interesting relationship between the optimum tariff and the tariff rate that would serve to maximize the tariff revenue. Assume that an increase in country 1's tariff rate would be protective in the sense of raising p.[3] In the neighborhood of a tariff that maximized tariff revenue (i.e., one for which $d\{(T_1-1)\beta M_1\}$ were zero) an increase in the tariff would lower real income. That is, a revenue maximizing tariff rate is higher than the optimum rate.

The domestic price of imports, p, equals

[1] Consider a utility function for country 1 of the form $U_1 = U_1(A_1, B_1)$. With differentiability, no satiation, and both commodities consumed, dU_1/U_{1A} equals $dA_1 + (U_{1B}/U_{1A})dB_1$. In a competitive world p equals the marginal rate of substitution, U_{1B}/U_{1A}. Now define dy_1 as dU_1/U_{1A}, which is free of cardinal utility connotations.

[2] Note that the tariff is levied on the world price, β.

[3] It can be shown that this must be the case in the neighborhood of the optimum tariff as ϵ_2 must exceed unity.

T_1 times the world price ratio, β. Differentiating this relationship and substituting into (8) yields expression (9)

$$(9)\quad dy_1 = -M_1 d\beta + (T_1 - 1)\beta dM_1$$

as an alternative form for the change in country 1's real income. It illustrates the two basic means by which country 1's real income can be improved: first, by an improvement in 1's world terms of trade ($d\beta$ negative), which would serve to increase real income by an amount proportional to the original volume of imports, and second by any increase in imports in a situation in which the relative domestic valuation of the imported commodity, p, exceeds the relative cost of obtaining these imports on the world market, β. Note that $(T_1-1)\beta$ is precisely this price spread, $p-\beta$.[4]

With this expression for dy_1 at hand, it only remains to break down the separate price and income effects on the demand for imports. Differentiate (5) totally to obtain (10).

$$(10)\quad dM_1 = \frac{\partial B_1}{\partial p}dp + \frac{\partial B_1}{\partial y_1}dy_1 + \frac{db_1}{da_1}\frac{da_1}{d\left(\frac{1}{p}\right)}\frac{1}{p^2}dp$$

Note that db_1/da_1, the slope of the transformation schedule, is given by $-1/p$.[5] Further simplification involves the definition of elasticities to capture: first, the substitution effect of a price change on the consumption of the imported commodity, $\bar{\eta}_1$; second, the marginal propensity to import, m_1; and third, the substitution effect as production changes along the transformation curve, e_1. Following the procedure used in Jones [1], define:

$$\bar{\eta}_1 \equiv -\frac{p}{M_1}\frac{\partial B_1}{\partial p} \geq 0$$

$$m_1 \equiv p\frac{\partial B_1}{\partial y_1} \geq 0 \qquad \text{(by assuming no inferior goods)}$$

$$e_1 \equiv \frac{\left(\frac{1}{p}\right)}{(a_1 - A_1)}\frac{da_1}{d\left(\frac{1}{p}\right)} \geq 0$$

These relationships can be substituted into (10). Finally, express dp in terms of $d\beta$ and dT_1 to obtain (11) as the expression for the relative

$$(11)\quad \hat{M}_1 = \frac{-\left[\bar{\eta}_1 + \frac{m_1}{T_1} + \frac{e_1}{T_1}\right]\hat{\beta} \quad -\left[\bar{\eta}_1 + \frac{e_1}{T_1}\right]\hat{T}_1}{\left[1 - m_1\frac{(T_1-1)}{T_1}\right]}$$

change in country 1's imports. Comparing this with the original breakdown of $\hat{M}_1$ in (3) it is possible to separate out the expressions for the elasticity of country 1's offer curve, ϵ_1, from the expression for the shift in its offer curve, given by $-\bar{\epsilon}_1$. Thus:

$$\epsilon_1 = \frac{\bar{\eta}_1 + \frac{m_1}{T_1} + \frac{e_1}{T_1}}{1 - m_1\frac{(T_1-1)}{T_1}}$$

$$\bar{\epsilon}_1 = \frac{1}{\left\{1 - m_1\frac{(T_1-1)}{T_1}\right\}}\left\{\bar{\eta}_1 + \frac{e_1}{T_1}\right\}$$

Several comments are in order. First, note that the expression in the denominator is reminiscent of a Keynesian type of multiplier phenomenon. If initially there is free trade, this multiplier effect disappears. If $T_1 > 1$, any increase in import spending increases tariff revenue. This increase in income is, in part, respent on imports, which

[4] For a discussion of the sources of a change in real income in a more complicated model involving capital mobility see Jones [2].

[5] If country 1 is specialized in producing commodity A, without there being a corner tangency, $db_1=0$ and the last term in (10) disappears.

serves further to increase tariff revenue in the next round. It is this phenomenon which makes it difficult to draw a tariff-ridden offer curve in a simple step-by-step fashion. Secondly, note that the numerator of ϵ_1 draws together two substitution terms ($\bar{\eta}_1$ in consumption and e_1/T_1 in production) which are positive, and a positive income effect. Finally, observe that the numerator of $\bar{\epsilon}_1$ consists only of the substitution terms. This is what moved me to call the shift expression $\bar{\epsilon}_1$, after using $\bar{\eta}_1$ to denote just the substitution effect in consumption. The reason why this contains only substitution terms is most readily seen in the case of initial free trade, when $\bar{\epsilon}_1$ is $(\bar{\eta}_1+e_1)$. The term $-\bar{\epsilon}_1$ is the coefficient of $\hat{T}_1$. At initial free trade, real income can be affected by an increase in the tariff only if β, the world terms of trade, is affected and dy_1 would then equal $-M_1 d\beta$. Therefore, holding β constant (to capture the shift in the offer curve), one finds that an increase in T_1 reduces imports only by making commodity B at home more expensive and thus stimulating a reduction in consumption and an expansion in production (both substitution effects). If a tariff were initially in effect, real income would indeed be altered with β given, but as expression (9) shows, this real income change depends itself on a change in imports. It thus enters only via the multiplier phenomenon shown by the denominator in $\bar{\epsilon}_1$.

The expressions for the elasticity of country 2's offer curve, ϵ_2, and the shift in that offer curve induced by a change in T_2, $-\bar{\epsilon}_2$, being completely analogous, are not shown explicitly.

The analysis to this point has only involved changes in one country. Tariff theory is primarily concerned with the effects of tariff changes on world terms of trade, domestic prices, and real incomes. Answers to these questions can now easily be obtained by differentiating the market equilibrium condition shown in (12).

$$\beta M_1 = M_2 \tag{12}$$

At world prices the value of imports into each country must be equal. (Equilibrium is obtained at the intersection of the offer curves.) The relationship in terms of rates of change is shown by (13),

$$\hat{\beta} + \hat{M}_1 = \hat{M}_2 \tag{13}$$

and the solution for the change in the world terms of trade is found by substituting in (3) and (4) for the $\hat{M}_i$. This solution is exhibited in (14):

$$\hat{\beta} = \frac{1}{\Delta}\{\bar{\epsilon}_2\hat{T}_2 - \bar{\epsilon}_1\hat{T}_1\} \tag{14}$$

$$\text{where } \Delta \equiv \epsilon_1 + \epsilon_2 - 1$$

The denominator, of course, is positive by the Marshall-Lerner conditions for stability. Once it has been established that the $\bar{\epsilon}_i$ are positive, it is easy to see that an increase in country i's import duty must serve to improve its world terms of trade (unless one of the ϵ_i is infinite). This is a standard result, not challenged by Södersten and Vind.

The crucial relationship that prompts this analysis concerns the impact of tariff changes on the relative domestic price of imports in country 1. Since $\hat{p}$ equals $\hat{\beta}+\hat{T}_1$, easy substitution into (14) yields this price change in (15).

$$\hat{p} = \frac{1}{\Delta}\{(\Delta - \bar{\epsilon}_1)\hat{T}_1 + \bar{\epsilon}_2\hat{T}_2\} \tag{15}$$

The question is whether, as Metzler maintained, an increase in T_1 can serve to lower p. To answer this question, it is necessary to make use of the breakdown of ϵ_1 (for it is contained in Δ) and $\bar{\epsilon}_1$ already provided. Thus a tariff can fail to be protective if the inequality shown by (16) is satisfied:

$$\epsilon_2 + \frac{m_1}{T_1\left\{1 - m_1\dfrac{(T_1 - 1)}{T_1}\right\}} < 1 \tag{16}$$

Sufficiently small values for ϵ_2 allow this possibility.

Since this point has been challenged so strongly by Södersten and Vind, it is worthwhile to attempt a simple verbal explanation.

This is most easily done in the case of initial free trade, in which case a tariff levied by country 1 fails to be protective if $\epsilon_2 < 1 - m_1$. Consider the world market for 1's export commodity, A. In a stable market an increase in T_1 would raise A's domestic price in country 1 (and thus lower p) only if excess demand for A would be created at the initial value of the domestic price ratio, p. In country 1, the increase in T_1 would serve to lower β, i.e., improve 1's world terms of trade and therefore real income, at the initial domestic price ratio. Part of this increase in real income, $1 - m_1$, would spill over to increase 1's demand for its export commodity. This is the only source of change in demand and supply in country 1 at initial p since substitution terms do not operate. Opposed to this increased demand for A in country 1 is the reduction in net demand in country 2 as the decrease in β (at initial p) moves country 2 back along its offer curve for imports of A. (Note that 2's offer curve does not shift as, by assumption, it keeps T_2 constant at unity.) If the increase in demand for A by the home country, $1 - m_1$, exceeds the reduction in net demand abroad, given by ϵ_2, excess demand for A at initial p is created, forcing p down.

To conclude this section, consider the impact of tariff changes on country 1's real income. For this purpose, it is convenient to substitute $dM_2 - M_1 d\beta$, for the βdM_1 term found in (9). Then use (4) for dM_2 to obtain (17) as the expression for dy_1.

$$dy_1 = -\beta M_1 \{ [(T_1 - 1)(1 - \epsilon_2) + 1]\hat{\beta} + (T_1 - 1)\bar{\epsilon}_2 \hat{T}_2 \} \tag{17}$$

From this the formula for country 1's optimal tariff is obtained by noting that the rate must be such that the coefficient of $\hat{\beta}$ is zero. This immediately gives the optimal tariff formula shown in (18). Of course (18) presupposes that ϵ_2 exceeds unity.

$$t_1^{opt} = \frac{1}{\epsilon_2 - 1} \tag{18}$$

Otherwise, as (17) shows, an increase in the tariff rate by country 1 would, by reducing β, serve always to increase 1's real income. Although not necessary in deriving the formula for the optimum tariff, the solution for $\hat{\beta}$ in terms of the $\hat{T}_i$ shown by (14) can be substituted into (17) to obtain expression (19) for dy_1.

$$dy_1 = \frac{\beta M_1}{\Delta} \{ \bar{\epsilon}_1 [(T_1 - 1)(1 - \epsilon_2) + 1]\hat{T}_1 - \bar{\epsilon}_2 [(T_1 - 1)\epsilon_1 + 1]\hat{T}_2 \} \tag{19}$$

The extra bit of information this yields is that any increase in 2's tariff rate must (by raising β) worsen country 1's real income.

II. *The Error in the Södersten-Vind Analysis*

The formal analysis presented by Södersten and Vind is not easy to follow and contains some minor errors that are not crucial.[6] The vital error is contained in the assumption made in [5], footnote 2 on p. 401. In my terminology, they are assuming that an increase in the relative domestic price of A in country i always raises i's consumption of B when income is constant in B-units. Most importantly, as I shall demonstrate below, this assumption is far from innocent, for it implies that for country 2, ϵ_2 must exceed

[6] As an example of the minor errors, according to my calculations, the coefficient of dt_1 in their equation (24), (p. 400) should read a_2 instead of a_1. Similarly, in equation (25) (p. 400) the last term in the coefficient of dt_1 should be $C_{1s}(\partial C_{1m}/\partial Y_1)$ instead of $S_{1x}(\partial C_{1m}/\partial Y_1)$, and the factor $(1+t_2)$ should be deleted from in front of the bracketed expression for β_2 in equation (19) (p. 399) and in between the two bracketed expressions in the coefficients for dt_2 in equation (25) (p. 400). Aside from the complexity inherent in treating this problem formally as an eleven equation model, their instructions how to proceed in their model are occasionally misleading. For example, equations (1a)–(11a) (p. 397) purport to represent differentials of equations (1)–(11) (pp. 396, 397). However, equations (6a) and (7a) merely restate equations (8) and (9). Similarly, equation (3a′) (p. 398) follows not from equation (3a) (p. 397), but from (4a), with substitutions from equations (8a)–(11a) (p. 397). As a final example, equation (3a″) (p. 399) is not a rewritten form of (3a′). Rather, it follows from the balance of payments equilibrium condition with the aid of equations (1a), (2a), (6a), and (7a), (p. 397). As I say in the text, these errors are not crucial, but they do make it unnecessarily difficult to work through their model.

unity. And, as we saw in condition (16), this is sufficient to rule out the Metzler effect.

Before proceeding to the proof, however, note that the model I have developed in the preceding section incorporates the same assumption about the distribution of the tariff proceeds as does the Södersten-Vind model: all proceeds are distributed to the private sector in lump-sum fashion. I stress this because Södersten and Vind claim that the difference between their results and the Metzler result derives from Metzler's assumption that the government spends the tariff proceeds according to its own arbitrary taste pattern. Furthermore, they suggest that Metzler's different result may in part be attributed to the fact that the foreign country was assumed not to levy a tariff.[7] On both counts the analysis presented in the preceding section, where Metzler's results are confirmed, adopts the same assumptions as Södersten and Vind.

The point I wish to make is perhaps best illustrated in Figure 1. To simplify, I assume that production in country 2 is represented by the endowment point, E, and is unresponsive to price changes.[8] Let the initial budget line be represented by the solid line QEN, and the initial consumption bundle by point ①. If π increases, the budget line rotates around E, becoming flatter. Suppose, instead, that at this new value for π income had been increased so that a new budget line, QV, preserves the same income in B units (OQ) as before the price change. The Södersten-Vind assumption is that the consumption bundle on budget line QV, point ②, must lie below ①. That is, the higher price for B must reduce 2's consumption of B. However, since the income effect (consume more B) runs counter to the substitution effect (consume less B) when π rises (and income in B-units remains at OQ), it would not be unusual for the consumption point to have been northeast of ①. In any case, the

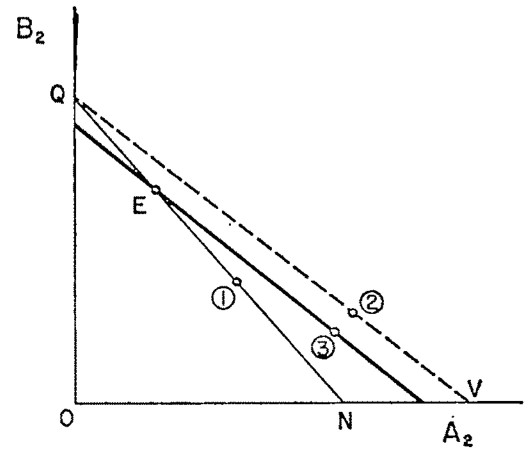


FIGURE 1

fact that the correct new budget line must pass through the production point, E, implies that if no goods are inferior the new consumption point, ③, involves an even lower consumption of B. Country 2's offer curve (not drawn) passes through E and points ① and ③ and therefore must be elastic. Södersten and Vind have implicitly assumed that $\epsilon_2 > 1$, and therefore by assumption have ruled out the possibility of Metzler's result.[9]

In conclusion consider the expression for the optimal tariff displayed by Södersten and Vind in their equation (34) [5, p. 406]. They claim that this is a new formula and, indeed, in the guise presented this appears to be the case. However, substitution reveals this to be the same expression, $1/(\epsilon_2-1)$ displayed in my (18) for the optimal tariff, and this is a standard result.

REFERENCES

1. R. W. JONES, "Stability Conditions in International Trade: A General Equilibrium Analysis," *Internat. Econ. Rev.*, May 1961, *2*, 199–209.
2. ———, "International Capital Movements and the Theory of Tariffs and

[7] See [5], pp. 402–405.

[8] As the expression for ϵ_2 shows, the assumption I am making that $e_2=0$ serves to minimize the value of ϵ_2. I am demonstrating that this minimum value exceeds unity if the Södersten-Vind assumption that $\partial B_2/\partial(1/\pi)$ is positive is made.

[9] In equation (28) of [5], p. 401, the assertion is made that $(a_2\beta_1-\Delta)$ is negative. This is not necessarily so if $\partial C_{2\pi}/\partial P$ is negative. In my terminology it can be shown that $\partial B_2/\partial(1/\pi)$ with income in B-units held constant equals $\beta M_1\bar{\eta}_2-A_2(1-m_2)$. If this is assumed positive, substitution into the expression for ϵ_2 reveals that ϵ_2 must exceed unity.

Trade," *Quart. Journ. Econ.*, Feb, 1967, *81*, 399–431.
3. L. Metzler, "Tariffs, the Terms of Trade, and the Distribution of National Income," *Journ. Pol. Econ.*, Feb., 1949, 57, 1–29.
4. ———, "Tariffs, International Demand and Domestic Prices," *Journ. Pol. Econ.*, Aug., 1949, *57*, 345–51.
5. B. Södersten and K. Vind, "Tariffs and Trade in General Equilibrium," *Am. Econ. Rev.*, June, 1968, *58*, 394–408.

TARIFFS AND TRADE IN GENERAL EQUILIBRIUM: REPLY

By Bo Södersten and Karl Vind*

We begin this rebuttal with a few comments on what Ronald Jones calls "minor errors" in our paper [1]. We then turn to the main issue and demonstrate the principal requirements for ensuring a certain result for the effect of a tariff on domestic prices. Finally, we comment on the formula for the optimum tariff which we have derived, and use this example to illustrate the superiority of the explicit kind of analysis we have introduced for treating tariff theory.

First, the "minor errors": Jones' first three remarks in footnote 6 concerning coefficients are correct: the coefficient for dt_1 should read a_2, not a_1 etc. His further and more substantial remarks on the structure of the model, about equations (1a)–(11a), however, are incorrect and makes it somewhat doubtful whether Jones has really understood the nature of the model. It is wrong to think that equations (6a) and (7a) are merely a restatement of (8) and (9); they do not even contain the same variables. (6a) and (7a) are linear equations with the unknowns dS_{1m}, dS_{1x}, dS_{2x}, and dS_{2m} while (8) and (9) are nonlinear equations with unknowns P, S_{1x}, S_{2m}. The rest of Jones' remarks in the footnote are irrelevant. We have not maintained that (3a) follows directly from (3) or that (3a″) is a direct rewriting of (3a′). These transformations are made by using information implicit in the equation system.

Let us now discuss the main issue. It should be pointed out that Jones uses the same type of model as we do. The difference is that Jones starts out from equations for the offer curves, whereas we go directly to the basic underlying supply and demand functions. Jones claims that our model is "unnecessarily cumbersome." We maintain that the simplicity of his technique is misleading. To begin with, it only gives him results in terms of the functions for the offer curves, which are not very interesting. He must then go on and try to find the results in terms of more basic functions, which are never explicitly stated. His is a "Chinese boxes" technique which tends to obscure the real goal of the analysis. His refusal to start the analysis from an explicit formulation of the problem also makes it difficult for him to see the nature of the assumptions needed, and his perspective becomes at times quite distorted. Lack of space precludes us from repeating the definitions of the variables involved; we refer the reader, instead, to our original paper.

Jones maintains that our "vital error" is contained in an assumption about two partial derivatives, $\partial C_{1m}/\partial P$ and $\partial C_{2x}/\partial P$ being positive. First, we make the elementary observation that an assumption can be neither right nor wrong. Jones' assertion that our assumption contains a "vital error" is therefore nonsensical.

What is open to argument is the relevance and reasonableness of the assumptions. This we now turn to and interpret Jones' critique in that fashion. The assumption we used was that b_1 and b_2 are positive [1].[1] This assumption is sufficient for our results. It can be argued that it is too strong an assumption. We shall now give the necessary assumptions needed to achieve the results.

* The authors are, respectively, docent in international economics at the University of Lund, Sweden, and acting professor of economics at the University of Copenhagen, Denmark.

[1] The assumption stated in [1, footnote 2] implies b_1 and b_2 positive, but is not necessary for the results.

There are two issues at stake. First, we want to demonstrate which assumptions are needed to make the determinant, Δ, positive; second, we want to demonstrate which assumptions are needed to make the coefficient of dt_1 in expression (28) in our original paper [1] negative. A positive determinant is necessary in order not to get a reversal of all the main results. To have the coefficient in (28) negative is necessary for the effect of a tariff on the home market price ratio to be unambiguous. We will demonstrate that these two issues are intimately linked together.

In analogy with the theory of one consumer it is assumed that β_1 and β_2 are non-negative [1, p. 399]. As b_1 is larger than β_1 it also follows that b_1 is non-negative. Furthermore, we assume that a_1 and a_2 are non-negative (this is also assumed by Jones) [1, p. 399]. We have, in our paper, moreover, assumed that b_2 is non-negative.

The determinant, Δ, equals $a_1b_2+a_2b_1$ [1, p. 400]. From this, it follows that $\Delta>0$ is equivalent to $a_1b_2/a_2b_1>-1$. For the expression (28) to be negative, it is necessary that $\Delta-a_2\beta_1>0$ [1, p. 401]. From this it follows that $\Delta-a_2\beta_1$ being positive is equivalent to $a_1b_2/a_2b_1>-1+\beta_1/b_1$.

Now we know that $0<\beta_1/b_1<1$. Therefore, it is somewhat stronger to assume that $a_1b_2/a_2b_1>-1+\beta_1/b_1$, than to assume that $a_1b_2/a_2b_1>-1$. But essentially the same type of assumption is involved in both cases. In both instances, the left-hand side of the inequality can only be larger than the right-hand side provided that either $\partial C_{1m}/\partial P$ or $\partial C_{2x}/\partial P$ (or both) are negative.

Jones is correct to point out that there is a small interval where $a_1b_2/a_2b_1>-1+\beta_1/b_1$ without the determinant being negative. Let us, in memory of Metzler's two early pioneering articles, call this interval the Metzler interval. It might be useful to illustrate this interval in the following way:

The part of the line with short dashes is the "Metzler interval." Now two comments are in order.

First, it is a necessary and sufficient condition, for a_1b_2/a_2b_1 to be in the "Metzler interval" for the domestic price of importables to fall because of an increase in tariffs. It is, however, fully possible for the normal result to come about even though the value of b_2 should be negative.[2]

The second and more important comment concerns Jones' formulation of the problem. He is concerned only with what we have termed the "Metzler interval" and seems to be unaware of the true nature of the issue at stake. He airily assumes that a_1b_2/a_2b_1 can never be less than -1, but takes great pains with the case where it is slightly larger than -1.[3] With reference to Figure 1, it means that he is greatly worried about the small dashed area between -1 and 0, but pays no heed to the much larger area to the left of -1. This can hardly be characterized other than as an example of blind guides who strain at a gnat and swallow a camel.

At the end of his comment, Jones asserts that the formula for the optimum tariff which we have derived is not a new result, but the same expression as displayed in his formula for the optimum tariff.

First, we can note that Jones' statement is merely an assertion, as long as he has not

[2] Our speaking about a "Metzler interval" should not make the reader believe that Metzler's early analysis was faultless. Metzler, for instance, never explicitly brought supply factors into the analysis and only reasoned in terms of marginal propensities to consume and demand elasticities as the critical factors for getting the perverse result for home market prices. His treatment of k, the government's marginal propensity to consume, as a constant was also peculiar [1, p. 404].

[3] Jones asserts that the denominator is positive "by the Marshall-Lerner conditions for stability." No such analogue of a dynamic kind can be valid here. The only way Jones can get the determinant positive is by assuming a_1b_2/a_2b_1 never to be less than -1.

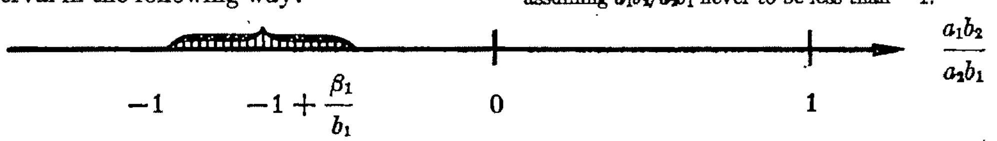


FIGURE 1

explicitly set out the elasticities etc. which make up his formula. The main point, however, is of another kind.

The message of our paper was, that by setting out a model where the basic functions are explicitly stated, we were able to derive results easy to interpret. These results were formulated in terms of the characteristics of the basic functions, and there was no ambiguity left in interpreting the variables making up the formulas.

Jones now states that by using a slightly different version of the model we used, he will get the same results as we did. Provided that the calculations are made correctly, this is what will happen; this should not astonish him unduly. He then tries to relate these results to earlier results formulated in terms of offer curves, maintaining that the new results are not new, but really implied in the older ones.

This is all well and good. In his eagerness, however, Jones loses again his sense of perspective. He forgets that the formula for the optimum tariff which we have derived contains much more information than the older one, because it is formulated in explicit terms. This is precisely what we wanted to demonstrate in our paper.

REFERENCE

1. Bo Södersten and Karl Vind, "Tariffs and Trade in General Equilibrium," *Am. Econ. Rev.*, June 1968, *58*, 394–408.

EXCHANGE RATES, CAPITAL FLOWS AND MONETARY POLICY

By E. Ray Canterbery*

Debates on foreign exchange models, statistical tests of short-term capital flows, and implications for monetary policy have recently abounded in this *Review*.[1] Now a new statistical test by Hans Stoll [15, 16] focuses upon some important concepts overlooked by Jerome Stein's detractors [6]. Stoll's discussion also illustrates some of the problems in analyzing foreign exchange under partial equilibrium conditions, particularly when monetary policy conclusions are derived therefrom. There continues to be confusion in the literature between what is valid in theory and what is directly testable statistically, and between the forces that are endogenous and those that are exogenous.

To my knowledge there are two full equilibrium models of simultaneous spot and forward foreign exchange market transactions in print: those of Stein [12] and Canterbery [5]. Both authors later modified their models by explicitly specifying the speculative variable as endogenous rather than exogenous [14], [4]. As I owe a great debt to Stein's pioneering work [13], it is not surprising that my article [4, p. 32] derives a forward premium solution (though via more elaborate equation specifications) similar to Stein's [12, p. 56].[2]

I. *Interest Parity and Speculation*

Recent empirical studies can be interpreted in terms of both the behavioral equations and the interest parity solutions of the above theoretical models. Using the foreign exchange supply and demand equations specified by Canterbery [4], a simultaneous solution will yield equilibrium spot and forward rates. Given these rates, a new interest rate parity requirement is derived [4, p. 409, equation (24)]. Using Stoll's symbols, S = spot foreign exchange rate; F = forward foreign exchange rate; S^e = spot rate expected by speculators; r^d = U.S. interest rate; r^f = foreign interest rate; F^* = forward rate satisfying traditional interest parity

* The author is visiting associate professor of economics at Simon Fraser University and associate professor at the Florida State University. He expresses appreciation to J. Lawrence Hexter, Jerry Shipley and Paul Meyer for some very useful discussions on econometric estimation problems in foreign exchange. They share no blame for possible errors.

[1] Five "Comments" [6] were incited by an article by Jerome L. Stein [12].

[2] There are two main differences in the solutions: Stein includes a Central Bank intervention variable and also calls the net speculative variable "exogenous" in [12] even though it is endogenous. The reader is referred to [4, pp. 398–99] for a listing of other differences.

(IRPT), so $F = S - (\hat{r}^f - \hat{r}^d) + f(\Delta S^e)$.[3] Assuming identical forward rate and Treasury bill maturities, the percentage forward premium in the modern theory (MT) solution is

$$(1) \quad (F - S)/S = -(r^f - r^d) + f(\Delta S^e)/S.$$

There is no unique value for $(F-S)/S$ as the exchange rates are, in part, determined by a speculative variable. This MT solution implies that all variance unexplained by the interest differential can be explained by a speculative variable. This is a hypothesis crucial to the acceptance of recent empirical parity tests, and adds theoretical credibility to a number of statistical findings suggesting that interest differentials under narrow gold bands do not entirely explain forward premiums [e.g., 7, 1, 2]. Traditional interest rate parity would be

$$(2) \qquad (F - S)/S = -(r^f - r^d).$$

Stein does call attention to the interest rate parity implications of his solution in [11], and Peter B. Kenen [8] has emphasized the interest rate parity aspects of foreign exchange in a single-firm model. Kenen demonstrates definitionally what the present author has proven with behavioral relations: namely, traditional interest rate parity prevails only when speculation is nonexistent or stabilizing, that is, when the speculative coefficient in the above parity form solution (1) is zero or near-zero [8, p. 150]. Stoll's and Stein's statistical analyses could be viewed as either tests of the Stein-Canterbery solution or of Kenen's hypothesis.

Stein's final solution for the forward premium (though he confuses readers at various points by calling the forward premium the forward rate) is the beginning of his statistical test. Stein isolates, in theory, the speculative variable with his MT solution. Because of the way he specifies his model, however, the speculative variable netted between the spot and forward exchange markets must be endogenous rather than exogenous as Stein claims. (In the model speculative capital flows and both the forward and spot rates interact.) The statistical problem is to identify correctly the precise nature of the speculative variable and design an appropriate test.

Using a linear form of equation (2) Stein regresses the absolute forward premium $(F-S)$ on the uncovered interest rate differential (r^d-r^f) and calculates the residual, the actual minus the predicted rate [12, p. 58]. Assuming speculation to be independent of interest rates, Stein presumes that this residual is an unbiased estimate of the speculative parameter. As Laffer [9, pp. 557–58] suggests, however, it is not possible to derive a statistically unbiased estimate of the parameter for an endogenous variable with this particular technique. A test of the new parity solution carries all the statistical problems associated with simultaneity.[4] In theory, nonetheless, the variable in this general solution is the net speculative function under flexible exchange rates or under a narrow gold band (away from the limits).

An overlooked significance of this new MT–$IRPT$ solution is its isolation of two variables (the interest differential and the speculative) from real trade flows. Stein, in his *Reply*, unnecessarily accepts the argument by Heckerman [6, pp. 555–56] and Laffer [9, p. 557] that export–import variations can alter significantly the difference between the forward and spot rates (forward premium). In the Canterbery model, trade can affect the premium only insofar as exchange rate expectations affect hedging and this impact may be very marginal. Even when traders hedge to avert all risk, they are reacting to the same variable as are speculators, namely the expected future exchange rate (which can at times be constant). Assuming hedging then to be a function also of S^e, it is subsumed in the speculative coefficient of the forward market anyway. The relation between spot and forward

[3] The caret (^) signifies that values originally defined as percentages are now in absolute terms. In this modern theory (MT) solution the interest differential is the unitary exchange value of extra return from investing abroad rather than at home for the period of the contracts. See [4, p. 409n] for the precise calculation.

[4] If the estimated equation were $(F-S)/S = \alpha_4 + \beta_4 R + u$, the residual ($u$), in fact, is $[\beta_5 S^e + v]$, where v is an error term and S^e is not independent of $(F-S)/S$. Herein R equals $(r^f - r^d)$.

exchange rates is otherwise unaltered by the trade balance (exogenously given). The only other possible source of variation in the net S^e is a change in the interest rate, a second source of statistical bias in the coefficients.[5] The variation in the forward premium unexplained by interest rates must be then explained by some speculative variable. If the net speculative coefficient can be estimated as an endogenous variable and the speculative interest rate interaction is not present (or else the statistical bias from interaction minimized), there should be little objection to using speculative observations calculated from this indirect parameter estimate as a proxy independent variable along with interest differential observations explaining short-term capital flows.[6]

As Heckerman and Laffer advocate, a trade variable is missing from Stein's capital flows equation. But the suggested unlagged trade surplus variable is dependent under flexible exchange rates. It is, moreover, an accounting near identity with short-term capital flows (as Stein correctly replies) under either flexible or fixed rates. Given these considerations and having identified all the major independent variables, specifying trade as a distributed lag variable may yield an improved estimating equation for capital flows.

II. *Empirical Tests and Policy Implications*

A speculative variable that is correctly specified for all eventualities is needed; whether speculation interacts with exchange rates, is correlated with interest rates, interacts with interest rates, or is unrelated to interest rates. This would merge theory and empirical verification. In an inquiry "... only into the relationship of the forward rate to the spot rate" [15, p. 61], Stoll performs an improved test. The speculative variable is specified as a function of lagged spot rates, making it endogenous and of a form similar to that suggested (in theory) by Canterbery [4, p. 404n] and also tested elsewhere by Stein and Tower [14] and Sven Arndt [1].[7] Stoll's dependent variable is F/S rather than $(F-S)$ and his speculative variable is S^e rather than ΔS^e.

Arndt's distributed lag specification for the expected exchange-rate may lead ultimately to improved speculative parameter estimates.[8] But Arndt's estimates of the covered (and uncovered) interest rate and speculative parameters against capital flows also are biased, a bias (as Arndt correctly explains) resulting from the application of non-simultaneous techniques to what is (in theory) a simultaneous system. Stoll focuses upon a part of the Stein-Arndt type of statistical problem. Stoll specifies his endogenous speculative variable in such a way as to reduce interest rate coefficient bias. As Stoll wisely recognizes that past expectations can influence current expectations (unless random), the most general expression of his behavioral equations would have a as the coefficient of $(S-S^e_{-1})$ and b as the coefficient of S^e (of current expectations). His theoretical equation is reducible to

$$(3) \quad F/S = c + (1-b)R + b(1-a)S^e_{-1}/S,$$

where R is the interest rate differential. As S^e and S^e_{-1} are unknowns, however, they are defined functionally in terms of knowns $(S, F, F_{-1}, R, R_{-1})$ and hence eliminated from the final estimating equation. The statistical equation is then fitted in the con-

[5] A hypothesis of independence is a true one only if the monetary authorities do not change interest rates in reaction to exchange speculation. Even if speculator's responses to such interest rate changes are random the statistical collinearity is already present though of a sign opposite to what it would be with the reactions hoped for by the authorities.

[6] The statistical independence that would cause the two variables to be orthogonal simply means that the addition to the multiple regression (against capital flows) of the speculative variable would not alter the interest rate coefficient and one has the long sought statistical ideal coinciding with a theoretical solution. The coefficient of multiple determination also would be higher. It is emphasized, however, that as a reduced form estimate the speculative parameter in the *MT* solution is only a proxy for speculation in the capital-flows equation.

[7] Stoll's speculative variable is specified as $S^e = aS + (1-a)S^e_{-1}$, implying that S^e is a geometrically declining weighted average of past spot rates [15, p. 65]. A second order lag generated better results for the U.K.

[8] Data used in the test were for Canadian capital flows, 1952–60.

strained form:

$$(4)\quad F/S - (1-a)F_{-1}/S = 1 - b(1-a) + (1-b)\cdot[R - (1-a)R_{-1}] + z.$$

The coefficient $(1-a)$ is the weight given to prior expectations and $(1-b)$ is the interest rate coefficient.

Beyond pointing out the overlooked significance of Stein's preliminary solution and citing theoretically derived justification for Stoll's ingenious econometric test, some of the substantive discussion of Stoll in [16], wherein he apparently relies upon his model of [15], is challenged. Working through examples of capital-flow changes prompted by exogenous disturbances under both fixed and free exchange rate systems, Stoll attempts to illustrate that under certain conditions Keynes' "two interest rate" policy will fail and that even under flexible rates monetary policy is not independent of "international considerations." In two fixed rate cases his partial equilibrium analyses are based on changes in R that are exogenous. Throughout, all exchange rates are from U.S. point of view, i.e., the dollar cost of one foreign currency unit.

In Stoll's second fixed rate case between the U.S. and the U.K., the exogenous influence is a rise in the U.S. interest rate. Stoll states on the one hand, "... the spot rate cannot change and no reason exists for expectations about future spot rates to change, ..." [16, p. 895]. On the other hand, he says that spot funds "... will be covered forward at F so that [the U.S.] F is driven up until $S^e = S < F = F^*$" [16, p. 895]. Why would interest rate arbitragers cover if it is expected that the spot rate will remain unchanged $(S^e = \overline{S^e})$? Interest arbitrage becomes a function of $[(r^f - r^d) + \overline{S^e}]$ rather than of $[(r^f - r^d) + (F-S)/S]$.[9] Without the forthcoming cover, the forward rate would not appreciate and U.K. capital outflows would simply continue on a mostly uncovered basis. The U.K. "two interest rate" policy would fail, as Stoll suggests, but not because the U.K. forward rate was falling, implying a lower future spot rate. An often overlooked contributor to this failure is the presence of *uncovered* interest arbitrage under a system that guarantees minimal losses for those who base some of their exchange activity on S^e.

A third case is under Canadian floating rates where the exogenous change is a rise in U.S. interest rates. The expected spot rate apparently is again exogenous. Stoll argues that "... the burden of the initial adjustment will be on S, · · · , and not on F so that $F = F^* = S^e > S$" [16, p. 896]. To assert that the burden of change is on the spot rate is *now* to presume that all interest-rate arbitrage is uncovered and exports-imports *unhedged* despite the greater risk of exchange rate fluctuation in a freely fluctuating system. Because of the necessity for Canadians to cover under free rates, it is more probable that the U.S. forward rate on the Canadian dollar will be appreciating while the spot rate falters.[10] The movement to a new equilibrium is that described by Stein [13, p. 32–35]. Rather than the continuing spot depreciation described by Stoll, *equilibrium* is reached at some lower value of S. Interest rate parity (equation (2)) is also likely achieved in a short time period.

Whether the exchange rate moves far enough to cause "... Canadian exports [to] increase ... and inflationary pressure ..." [16, p. 896] depends upon the arbitrage and import elasticities. The conclusion that "... the Canadian authorities refuse[d] to allow r^f to fall" in 1959 demonstrates only that the Canadian monetary authorities did not act independently of international considerations and contributed to a domestic recession; it does not prove domestic monetary policy could not have been independent.[11]

[9] Stoll's model excludes this possibility by defining all hedgers as those who take a riskless position. Canterbery [3, 4, 5] defines uncovered interest arbitrage as, in part, a function of ΔS^e. Importers and arbitragers only attempt to eliminate risk if in fact they have formulated some expectation about future rate movements, including expected no change. These reactions nonetheless are captured statistically in Stoll's estimates because of the residual nature of the speculative variable in equation (1).

[10] Normal leading and lagging by importers and exporters would reinforce these rate movements.

[11] A careful observer of Canadian economic policies

In estimating the modern interest rate parity equation (1), there can be two sources of simultaneity bias. In Stein's test, $(F-S)=h(\Delta S^e)$ and $\Delta S^e=g(F-S)$. However, monetary policy makers sometime change interest rates in reaction to changes in S and/or F. Thus, R may be also a function of past values of S and F, and hence of S^e (and vice versa).

Stoll in [15] estimates equation (2) in order to derive an intentionally biased estimate of the interest rate parameter. By specifying S^e as a function of predetermined spot prices in testing equation (1), however, Stoll eliminates the speculative exchange-rate bias encountered by Stein. Because of the nature of correlation and Stoll's technique, however, one cannot positively determine the direction of cause and effect[12] between R and S^e. The Canadian interest rate coefficient increased from .80 to .96 (under flexible rates) after addition of the speculative variable to the regression; and the U.K. coefficient *fell* from .80 to .60. Stoll attributes these R coefficient changes to the elimination of *all* bias in the interest rate parameter estimates. The cause-effect Stoll presumes [15, pp. 76–77] between interest differentials and speculative behavior may be at times of a direction opposite that which he asserts and at times nonexistent (qua Stein's independence assumption).

The Canadian authorities raised interest rates and S^e rose. Did S^e rise because of the interest rate advance or because of natural market forces? Speculation may have been stabilizing under freely fluctuating exchange rates. Yeager [18], Rhomberg [10], and Stein [14], have concluded that speculation tended to raise the rate on the Canadian dollar during the flexible rate period when it was tending to fall.[13] The U.K. authorities increased interest rates and S^e fell. Did S^e fall in perverse reaction to the interest rate or does this simply suggest that under a narrow gold band higher interest rates do not necessarily attract capital? In terms of monetary policy Stoll's comparative regression results may simply demonstrate that policy makers' reaction to capital flows was to change interest rates. Furthermore, under free rates the interest changes may have been unnecessary and under fixed rates, fruitless.

The foregoing discussion is not intended to fault Stoll's statistical test per se. Indeed, if all statistical biases were reduced, Stoll's coefficients approximate the true parameters for the reduced form equation in the Stein and Canterbery models. Stoll may have accomplished the statistical test only attempted by Stein and never tried by the present author. And Stoll's evidence does lend considerable support to the new, more general form of interest rate parity.

What general policy conclusions emerge from this debate? First, the new formulation of the *IRPT*, both theoretical and empirical, raises doubts concerning the advisability of using interest rates directly to influence capital flows. If traditional interest rate parity prevails with free exchange rates

has concluded that "... the primary cause in the ensuing difficulties should be sought in Mr. Coyne's [head of the Central Bank] attitudes towards monetary policy, and not to the flexible exchange rate mechanism as such ...": See [17, p. 235].

[12] Let the form of the *MT* estimating equation (1) be $F/S=\alpha_0+\beta_1R+\beta_2S^e+v$, where v is the error term. If S^e is regressed against R, we have $S^e=\alpha_1+\beta_3R$. In turn, if the speculative variable is dropped for a test of the *IRPT* equation (2), the expected value of the estimated regression coefficient $(\hat{\beta}_1)$ is $E\hat{\beta}_1=\beta_1+\beta_2\beta_3$. In words, the estimated interest rate coefficient is biased by the slope coefficient of S^e regressed on R. Stoll is able, then to calculate an implied regression slope for S^e against R by comparing his estimated *MT* coefficients for R and S^e with his biased *IRPT* coefficient estimate for R. In Canada, $\beta_3=-4.0$ and in the U.K., $\beta_3=.5$. Even if the final *MT* estimates for β_1 and β_2 are unbiased, the direction of cause-effect between R and S^e can no more be inferred in this case than it can in any simple correlation between two variables, one arbitrarily defined as "dependent" and the other "independent." The failure to take this into account explains most of Stoll's subsequent interpretive errors. The reader, for the sake of accuracy, should bear in mind that Stoll's *MT* estimates for β_1 are derived from a constrained equation (4) rather than directly from (1). In the *IRPT* tests, $E\hat{\beta}_1=.80$ in both Canada and the U.K. In the *MT* test, the coefficient moves close to parity in flexible rate Canada.

[13] This possibility escapes Stoll as he claims that in the U.K. "... a fall in the interest differential produces pessimism about the course of the pound relative to the U.S. dollar, thus causing S^e to fall relative to S." [15, p. 76] But one could obtain similar statistical results if a speculative variable outweighed the effect on capital flows of interest-differential changes.

(speculation being stabilizing), no interest rate alterations may be required. When independent speculative variable changes outweigh the effects of interest rate alterations under fixed exchange rates, interest rate policy is less relevant for exchange stabilization purposes. With flexible rates, *smaller* (self-limiting) amounts of capital may move more *rapidly* in response to interest rate changes. Because of the low-risk profitability of *uncovered* interest arbitrage with a narrow gold band, however, large capital amounts may flow as interest rates change and it is virtually impossible to separate speculative from interest-rate forces.[14]

It is not possible then for the monetary authorities to predict with any certainty the impact of interest rate changes on short-term capital flows under a narrow gold band. The Canadian monetary authorities may have used interest rate changes to support the Canadian dollar under a system that does not require artificial bolstering (flexible rates) and the U.K. may have used a similar device for pound support under a system and conditions in which such a policy cannot succeed. The failure of floating exchange rates may rest in the lack of desire of the monetary authorities to follow independent monetary policies under *any* system rather than resting on the precept that speculation necessarily will be destabilizing.

[14] For a model describing speculator response to interest rate changes under alternative systems, see Canterbery [3]. Uncovered interest rate arbitrage demand has been expressed elsewhere [4, p. 407]; [5, p. 36] as, in part, a function of the expected future price of exchange. The absence of traditional parity under a narrow gold band does not mean that capital will never flow in reaction to interest rate changes; such capital flows at times will be multiplied by associated rising, falling or stable expected spot prices.

References

1. S. W. Arndt, "International Short-term Capital Movements: A Distributed Lag Model of Speculation in Foreign Exchange," *Econometrica*, Jan. 1968, *36*, 59–69.
2. S. W. Black, "Theory and Policy Analysis of Short-term Movements in the Balance of Payments," *Yale Economic Essays*, Spring 1968, *8*, 5–78.
3. E. R. Canterbery, "The Delayed Peg and the Adjustment Mechanism," A Paper before the Southern Economic Association Meetings, Nov. 1968.
4. ——— ,"A Dynamic Theory of Foreign Exchange," *Nat. Bank. Rev.*, June 1967, *5*, 397–413.
5. ———, *Foreign Exchange, Capital Flows and Monetary Policy*, Studies in Internat. Fin., Princeton 1965.
6. H. P. Gray, D. G. Heckerman, A. B. Laffer, P. H. Hendershott, and T. D. Willett, "International Short-term Capital Movements: Comments," *Am. Econ. Rev.*, June 1967, *57*, 548–70.
7. H. Grubel, "A Neglected Aspect of Forward Exchange Theory and Policy," *Jour. Fin.*, Sept. 1963, *18*, 537–48.
8. P. Kenen, "Trade, Speculation, and the Forward Exchange Rate," in R. E. Baldwin (ed.), *Trade, Growth and the Balance of Payments*, Chicago 1965, 143–69.
9. A. B. Laffer, "International Short-term Capital Movements: Comment III," *Am. Econ. Rev.*, June 1967, *58*, 556–60.
10. R. R. Rhomberg, "A Model of the Canadian Economy under Fixed and Fluctuating Rates," *Jour. Pol. Econ.*, Feb. 1964, *72*, 1–31.
11. J. L. Stein, "The Forward Rate and the Interest Parity," *Rev. Econ. Stud.*, Apr. 1965, *32*, 113–126.
12. ———, "International Short-term Capital Movements," *Am. Econ. Rev.*, Mar. 1965, *55*, 40–66.
13. ———, *The Nature and Efficiency of the Foreign Exchange Market*, Essays in Internat. Fin., Princeton 1962.
14. ———, and E. Tower, "The Short-Run Stability of the Foreign Exchange Market," *Rev. Econ. Stat.*, May 1967, *49*, 173–85.
15. H. Stoll, "An Empirical Study of the Forward Exchange Market under Fixed and Flexible Exchange Rate Systems," *Can. Jour. Econ.* Feb. 1968, *1*, 55–78.
16. ———, "A Note on Exchange Rate

Changes and Capital Flows," *Am. Econ. Rev.*, Sept. 1967, *57*, 894–97.
17. P. WONNACOTT, *The Canadian Dollar* 1948–1962, Toronto 1965.
18. L. B. YEAGER, International Monetary Relations, New York 1966, 423–40.

SOME NOTES ON THE ELASTICITY OF SUBSTITUTION

By FRANK S. T. HSIAO*

The first part of this paper demonstrates that the use of the concept of curvature is useful in checking some properties of the elasticity of substitution between two factors. In particular, it is shown that C. S. Soper's recent discussion [6] on the elasticity of substitution in a noneconomic region can be derived easily from our method. The second part illustrates the elasticity of substitution in a per capita output-per capita capital diagram. It may be thought of as a counterpart to Ronald Jones's diagram in [5] in the factor space.

I. *An Alternative Approach*

Let the production function be given by $Y=F(N, K)$, where F is properly differentiable but not necessarily homogeneous with respect to labor N and capital K. Then the elasticity of substitution is defined by

$$\sigma = (\omega/k)dk/d\omega \tag{1}$$

where $\omega = F_N/F_K$, $k=K/N$. If the isoquant is strictly convex to the origin, then at each point on it, we will have $\tan\beta=\omega$ and $\tan\theta=k$ with $0<\theta, \beta<\pi/2$.

Hence substituting in (1), we have

$$\begin{aligned}\sigma &= \frac{\tan\beta}{\tan\theta}\frac{d\tan\theta}{d\tan\beta}\\ &= \frac{\sin\beta\cos\beta}{\sin\theta\cos\theta}\frac{d\theta}{d\beta}\end{aligned} \tag{2}$$

* The author is an assistant professor of economics at the University of Colorado (Boulder). He is indebted to the referee for useful comments.

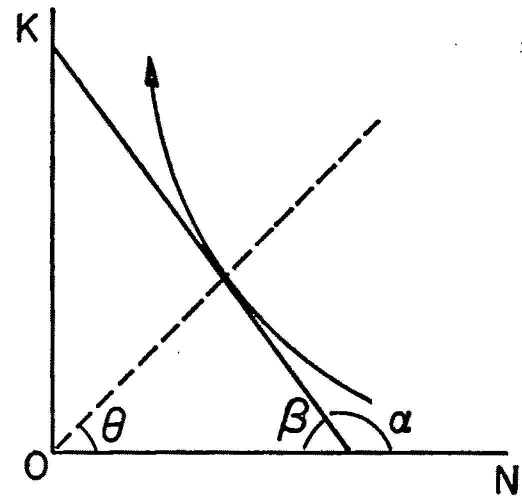


FIGURE 1

The merit of such a simple transformation lies in the fact that we can check immediately from (2) some well known properties of σ without complicated lengthy calculation, as has been shown in Allen [1, pp. 342–43].

From (2) we see immediately that in the usual economic region,[1] σ is positive and changes inversely with the curvature of the constant product curve in the range $0<\theta$, $\beta<\pi/2$: Since (2) can be written as

$$\sigma = \frac{\sin\beta\cos\beta}{\sin\theta\cos\theta}\left[\frac{d\theta}{ds}\Big/\frac{d\beta}{ds}\right] \tag{3}$$

where $\beta=\pi-\alpha$, and $d\beta/ds=-d\alpha/ds$, negative of the curvature, and s is the arc length[2] measured in the direction shown by the arrow from any fixed point P_0 on the iso-

[1] Economic region is defined here as the part of isoquant with negative slope ($\tan\beta>0$).

[2] Heuristically, by the Pythagorean theorem, the differential of arc length is, using the symbols of Figure 1, $ds=\sqrt{dN^2+dK^2}$, which may be integrated between appropriate limits to give the arc length

$$s=\int_{P_0}^{P}\sqrt{dN^2+dK^2}$$

see [7, p. 175]. The curvature of a curve is defined as $d\alpha/ds$, the sign of which depends on the effect of changes of s on the change of the angle α. Thus negative (or positive) curvature means that the concave side of the curve is directed to the right (or left) from a tangent whose direction coincides with increasing arc, as shown in Figure 1, Figure 2, and also part (I) of Figure 3. See [7, p. 404].

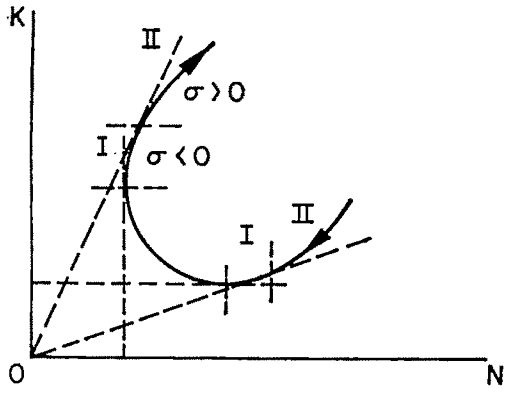


FIGURE 2

quant (see Figure 1). Thus from expression (3), each component is positive in $0<\theta$, $\beta<\pi/2$, and so σ is positive in this range. Notice that if the isoquant is concave to the origin, then the curvature will be positive, and so σ becomes negative.

For $0<\beta$, $\theta<\pi/2$, σ approaches infinity as the curvature $d\alpha/ds$ approaches 0, and σ approaches 0 as $d\alpha/ds$ approaches infinity. On the other hand, with finite curvature and $0<\theta<\pi/2$, $\sigma=0$ if $\beta=0$ or $\beta=\pi/2$, that is, if one of the marginal productivities of factors vanishes.

The elasticity of substitution of capital for labor and that of labor for capital are the same since, by (2), σ is symmetric. This may be more easily shown by writing (2) as

$$\frac{\sin 2\beta}{\sin 2\theta}\frac{d\theta}{d\beta}=\frac{\sin 2\left(\frac{\pi}{2}-\beta\right)}{\sin 2\left(\frac{\pi}{2}-\theta\right)}\frac{d\left(\frac{\pi}{2}-\theta\right)}{d\left(\frac{\pi}{2}-\beta\right)}$$

Finally, using our method, the sign of σ can be determined very easily in the uneconomic region,[3] which is defined as the part of the factor space where exactly one of the marginal productivities of factors are negative, thus $\tan\beta<0$.[4] If the isoquants are shaped like Figure 2, which is adopted by many textbooks, then the portion of isoquants in the uneconomic region can be divided into two parts (I and II in Figures 2 and 3). To determine the sign of σ in each part, we notice from (3) that

$$\begin{aligned}\text{sign }\sigma &= \text{sign }\{\beta'\cdot\theta'\cdot\cos\beta\}\\ &= -\text{ sign }\{\theta'\cdot\beta'\}\end{aligned}$$

where θ' and β' denote the derivatives with respect to s. Thus measuring the arc length s in the direction shown in Figure 1, and noting the simple effect of changes of s on the angles α, β and θ, the sign of σ in each part may be determined as follows:

	β'	θ'	σ
I	+	+	−
II	+	−	+

If the production function is linearly homogeneous, and the uneconomic region exists, then the isoquants have a shape like Figure 3, where a line through the origin will cut an isoquant only once [3].[5] The sign of σ in each part of the uneconomic region is the same as those in Figure 2, except that the reason for positive elasticity in region II is due to $\beta'<0$, $\theta'>0$.[6]

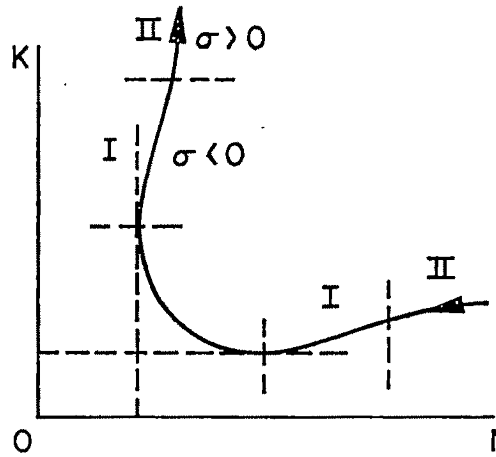


FIGURE 3

[3] Compare a similar attempt in Soper [6].

[4] See [3, pp. 300, 305], and also a discussion in [4, pp. 457–60].

[5] Solely for our purpose of illustrating the sign of σ, mathematical symmetry is preserved in Figures 2 and 3. For an economic argument against symmetry see, [3, pp. 305–07].

[6] Notice that in Figure 3, σ is discontinuous at the boundary between I and II: since $\beta'=0$ at the boundary point, σ approaches $-\infty$ as β' goes to 0 in I, and σ approaches $+\infty$ as β' goes to 0 in II.

Thus, σ in the uneconomic region is a measure of the elasticity of "complementarity": one of the marginal productivities is negative in this region, hence when the marginal rate of substitution decreases, both factors must be increased to stay on the same isoquant. But the relative rate of increase differs in Figure 2. In the upper part of Figure 2 where the marginal productivity of capital is negative, σ is negative in I because θ' is positive, which means that the demand for capital must increase faster than that of labor to stay on the same isoquant. This is so, since the capital-labor ratio k is less than the absolute value of the ratio of factor productivities, $|\omega|$, which is now measured by the angle α (less than $\pi/2$ in this case). On the other hand, in II, where $k>|\omega|$, σ is positive because θ' is negative, which means k must be decreasing so as to stay on the isoquant when (ω) is changed. Similar reasoning holds in the lower uneconomic region where the marginal productivity of labor is negative.

We note in passing that, as pointed out by Soper [6], when a production function has a positive elasticity of substitution in the entire factor space, then it can not have an uneconomic region.

II. *Two Illustrations*

Now assume that constant returns to scale hold in production, and in addition to labor and capital, technological change is incorporated into the production function through the time factor t, which operates so as to raise the output continuously over time: $Y=F(N, K; t)$. If perfect competition exists in the factor market, then wages (w) per unit of labor employed and profit (r) per unit of capital used will be equal to the (positive) marginal productivity of factors: $w=F_N$, $r=F_K$. Rewriting the production function into per capita magnitudes:

$$y = f(k, t) \tag{4}$$

where $y\equiv Y/N$, $k\equiv K/N$, $f(k, t)\equiv F(K/N, 1; t)$, we have the relation $r=f_k(k, t)$, and

$$w = f(k, t) - kf_k(k, t) \tag{5}$$

where w, as well as r, is a function of k and t. If in addition the law of diminishing returns prevails, then $f_{kk}(k, t)<0$, and so factor prices and capital-labor ratio have a unique relation.

Thus, from equations (4) and (5), we obtain

$$\frac{\partial y}{\partial w} = -f_k/kf_{kk}; \qquad \text{hence } \frac{w}{y}\frac{\partial y}{\partial w} = \sigma$$

(if $f_t\equiv 0$, then $\sigma=(w/y)(dy/dw)$, which is derived in Arrow, et al. [2]). From this expression of σ, the magnitude of σ in relation to unity can be shown in a per capita output-per capita capital diagram. In Figure 4, the curve OB' is the productivity curve for a fixed time point. At the capital-labor ratio OA, the output-labor ratio will be AB. The tangent line at B intersects with y-axis at W, which gives wages. On the other hand, when the capital-labor ratio is OA', the output-labor ratio is $A'B'$; the tangent line at B' intersects with the y-axis at W' and k-axis at P. Draw a supplemental line from P through W and intersect $B'A'$ at D; also draw a line BC parallel to k-axis.

Then

$$\Delta y/y = B'C/CA' = (B'A'/CA') - 1$$

$$\Delta w/w = W'W/WO = B'D/DA' = (B'A'/DA') - 1$$

hence

$$\sigma = \frac{\Delta y}{y} \Big/ \frac{\Delta w}{w} = \frac{(B'A'/CA') - 1}{(B'A'/DA') - 1}$$

therefore

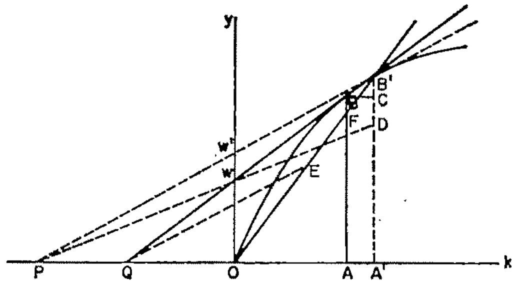


FIGURE 4

$$\sigma \gtreqless 1 \quad \text{according to} \quad CA' \lesseqgtr DA'.$$

Finally, we may use the original definition of σ given in (1) to express the magnitude of σ relative to unity in the y-k diagram.

The supplemental line QE is drawn parallel to the tangent line PB'. The ray OB' intersects the perpendicular line AB at F. Then, since tangent $\angle WQO = f_k$. $OQ = \omega$, $PQ = \Delta\omega$, we have

$$\begin{aligned} \Delta\omega/\omega &= PQ/QO = EB'/OE \\ &= (OB'/OE) - 1 \\ \Delta k/k &= AA'/OA = FB'/OF \\ &= (OB'/OF) - 1. \end{aligned}$$

Hence by equation (1),

$$\sigma = \frac{\Delta k}{k} \Big/ \frac{\Delta\omega}{\omega} = \frac{OB'/OF - 1}{OB'/OE - 1}.$$

Therefore

$$\sigma \gtreqless 1 \qquad \text{as } OF \lesseqgtr OE.$$

REFERENCES

1. R. G. D. ALLEN, *Mathematical Analysis for Economists*. New York 1938.
2. K. J. ARROW, H. B. CHENERY, B. MINHAS, AND R. M. SOLOW, "Capital Labor Substitution and Economic Efficiency," *Rev. Econ. Stat.*, Aug. 1961, *18*, 225–50.
3. G. H. BORTS AND E. J. MISHAN, "Exploring the 'Uneconomic Region' of the Production Function," *Rev. Econ. Stud.*, Oct. 1962, *29*, 300–312.
4. L. DE ALESSI, "The Short Run Revisited," *Am. Econ. Rev.*, June 1967, *57*, 450–61.
5. R. W. JONES, " 'Neutral' Technological Change and the Isoquant Map," *Am. Econ. Rev.*, Sept. 1965, *55*, 848–55.
6. C. S. SOPER, "The Elasticity of Substitution," *Econ. Record*, Dec. 1965, *41*, 539–48.
7. G. B. THOMAS, *Calculus and Analytic Geometry*. Reading. 1953.

PRODUCTIVITY GROWTH IN LATIN AMERICA: COMMENT

By CONSTANTINE MICHALOPOULOS*

For students of the progress made by Latin American countries in their uphill struggle for development, the findings of H. J. Bruton, in his recent article in this *Review*, might appear very disheartening [2]. Briefly, Bruton's main findings are as follows: first, that the mean rate of growth of total productivity of five Latin American countries between 1940 and 1960 was roughly half the productivity growth achieved by selected industrial countries over a comparable time period; second, that whatever productivity growth these Latin American countries did experience, particularly in the last decade considered, was not "pure" productivity change, but the result of increased capacity utilization [2, p. 1101].

The purpose of this note is to suggest that Bruton underestimated productivity growth in Latin America. The downward bias in Bruton's productivity estimates will be demonstrated through an analysis of the functional distribution of income in the Latin American countries examined.

I. *Relative Factor Shares*

Bruton's results are based on a model employing the Cobb-Douglas production function with disembodied technological change. He fitted such a function to data from selected Latin American and industrialized countries for the period 1940 to 1960. The central feature of his specification procedure should be noted: In the absence of direct estimates, the shares of capital and labor in total output (assumed to be constant over time) were used to approximate output elasticities. Total productivity growth was then calculated as an unexplained residual.

Bruton assigned a share to capital in Latin American countries close to double the share assigned to capital in industrial countries. He made no attempt to justify

* The author is associate professor of economics at Clark University.

this choice of shares or the assumption of their constancy over time. The shares used were explained by stating that "... they are consistent with a substantial body of evidence," [2, p. 1102] which, however, was neither reproduced nor cited anywhere in the study.

Of course, in determining the magnitude of productivity as a residual, the relative shares of the two factors used are of crucial importance. They act as weights for the rates of growth of inputs. The higher the weight assigned to the fastest growing factor, the lower is the calculated productivity [10]. On the other hand, if factor shares did change over the period in which they were assumed constant, the weights themselves would be biased. The weight assigned to the rate of growth of the factor whose share has increased would understate or overstate the proper weight, depending on whether it was closer to the actual share in the beginning of the period or at the end.

No effort was made to examine in detail the relative factor shares or the direction of change in these shares, if any, for the industrial countries examined by Bruton. However, some general observations appear warranted.

There is evidence that labor's share in the United States has been slowly increasing in recent periods [9]. Bruton uses a labor share for the United States equal to 75 percent, which can only be an end period share. On the other hand, the 70 percent labor share employed for other industrial countries is considerably higher than the labor share reported by Kerr [7, pp. 275–76] for several countries included in Bruton's sample.[1] In both cases the overstatement of the weight assigned to the slower growing input, labor, would tend to impart an upward bias to productivity estimates in the industrial countries group.

II. *An Empirical Investigation*

In light of the structural transformation undergone by the Latin American countries in the last twenty-five years, it might be reasonable to expect changes in relative factor shares, though it might be difficult to postulate on the basis of a priori considerations the direction of such changes. To determine the magnitude and the changes, if any, of relative factor shares in the Latin American countries, a separate empirical investigation was undertaken.

The investigation was limited to four of the five countries examined by Bruton: Argentina, Brazil, Chile, and Colombia. Mexico was excluded because data were not available in sufficient detail to allow estimation of factor shares. Data for different years were available for each country. However, in all four cases the data covered a span of at least eleven years.

In general, the investigation suggests that constancy in relative factor shares is a good first approximation for all four countries. On the other hand, the analysis amply demonstrated that the relative factor shares used by Bruton are inconsistent with the facts. In all four countries Bruton's weights for labor understate the actual weights and result in an understatement of total productivity change. Table 1 shows the evolution of labor's share of national income in the four countries over time.

The share of labor was computed by summing the payments made to labor in the form of wages and salaries including fringe benefits, and that portion of income originating in unincorporated enterprises that was a payment for labor services. Data on wages, salaries, and fringe benefits were available in all cases for the years examined. However, data on the share of income from unincorporated enterprises attributable to labor were available only for Chile. In the remaining three countries, this share had to be imputed.

No satisfactory method of imputing the labor component of income originating in unincorporated enterprises has been devised. One approach that may be used is to attempt to compute directly the share of one of the factors by multiplying its average rate of return times its employment in unincorporated enterprises and compute the share of the other factor as a residual, by subtracting from total income originating in the sec-

[1] Labor share in Sweden was 60.7 percent in 1951, in The Netherlands 48.3 percent in 1949, in France 46.0 percent between 1947 and 1950.

tor [8]. An alternative suggested by Denison [3, p. 256] is to assume that labor return forms the same portion of total income originating in unincorporated enterprises as it does in corporations.

Unfortunately, no data were available to make either of the above methods of computation possible. So it was decided to assume,

TABLE 1—LABOR'S SHARE OF NATIONAL INCOME IN FOUR LATIN AMERICAN COUNTRIES

(In Percent)

Years	Compensation of Employees (1)	Labor Payments in Income Originating in Unincorporated Enterprises (2)	Total Labor Share (1+2)
Argentina			
1950–1954	51.3	23.2	74.5
1954–1957	49.5	24.3	73.8
1957–1961	45.2	26.9	72.1
1950–1961	48.5	25.0	73.5
Bruton's Estimate			60.0
Brazil			
1950–1954	42.3	27.2	69.5
1954–1957	45.5	26.1	71.6
1957–1960	47.6	24.7	72.3
1950–1960	44.9	26.1	71.0
Bruton's Estimate			55.0
Chile			
1940–1945	43.0	23.6	66.6
1945–1950	43.5	23.2	66.7
1950–1954	45.7	24.0	69.7
1940–1954	44.0	23.6	67.6
Bruton's Estimate			50.0
Colombia			
1950–1954	38.9	29.9[a]	68.8
1954–1957	39.3	28.5	67.8
1957–1962	41.9	23.3	65.2
1950–1962	40.4	26.6	67.0
Bruton's Estimate			55.0

[a] For Colombia, income from unincorporated enterprises had to be imputed from data including property income. This was done by assuming that property income was 300% of the sum of corporate savings and direct government taxes on corporations. This was only a rough estimate arrived at by examining detailed data available for Brazil and Argentina. It is believed to be on the high side leading to an underestimate of the total labor share.

Sources: For Argentina [8], Brazil [9], Chile [10], Colombia [3].

as a tentative first approximation, that the portion of unincorporated income attributable to labor in the three countries for which data were not available was constant and equal to the average share of labor in unincorporated income in the United States for the years 1900 to 1949 as presented in Johnson [5, p. 178]. This average share turned out to be 63 percent. The assumption that the share has been constant is reasonable if the period is relatively short as it was in this case. The assumption that it was the same as in the United States in earlier periods is obviously open to objections relating to the comparability of the structure of production, the technologies employed, and relative factor scarcity in the two sets of countries. However, as noted, the assumption was only a tentative one and the intention was to discard it if the imputed estimates derived by employing the United States share diverged significantly from those calculated directly for Chile.

A comparison of the imputed labor component of unincorporated income in Argentina, Brazil, and Colombia, computed by using the U.S. share, to labor income originating in unincorporated enterprises in Chile, revealed striking similarities. These similarities among countries at similar stages of development strengthened confidence in the validity of what might be considered a drastic assumption, and enabled me to arrive at overall estimate of labor's share in national income.[2]

Table 1 indicates that the overall labor share in the four Latin American countries varied little over time. A coefficient of variation in the share of labor was computed by taking the ratio of the standard deviation to the mean labor share in each country. The computed coefficient ranged from a low of .015 for Argentina to a high of .029 for Colombia. By comparison the coefficient measuring variation in the share of employee compensation in the U.S. national income for the period 1940–55 was .032.

[2] The results would not be greatly affected if any other reasonable functional breakdown of income originating in the unincorporated sector is employed. If, for example, the factor weights used by Bruton are employed, the share of labor in national income would decline at most about 3 percent.

TABLE 2—GROWTH RATE OF OUTPUT (r_p), LABOR INPUT (r_l), CAPITAL INPUT (r_k), TOTAL PRODUCTIVITY (r_a)

(In Percent Per Year)

Country and Period	r_p (1)	r_a (2)	r_l (3)	r_k (4)	r_a/r_p (5)	r_a/r_p (6)
Argentina						
1946–1951	3.4	.6	2.4	3.9	.18	.12
1955–1959	1.7	−.3	1.5	3.4	−.18	−.35
1960–1964	1.2	0.	0.	4.6	−.02	−.50
Brazil						
1947–1953	5.6	2.2	2.4	5.9	.39	.29
1955–1959	5.6	2.1	2.8	5.2	.38	.30
1960–1963	5.0	1.5	2.8	5.1	.30	.24
Chile						
1940–1945	2.7	1.2	1.8	.9	.44	.52
1946–1953	3.9	1.5	2.1	3.0	.38	.36
1955–1959	3.0	.2	2.5	3.4	.07	.03
Colombia						
1946–1953	5.2	2.5	2.1	4.0	.48	.44
1955–1959	4.0	.6	2.6	4.9	.15	.10
1960–1964	4.5	1.7	2.0	4.3	.38	.33

Note: Columns (1), (3), (4), (6) are Bruton's estimates, [2, Table 1, p. 1103].
Source: See Table 1.

As shown in Table 1 in all four countries, the mean labor share was considerably higher than that used by Bruton. This in turn implies, as noted earlier, that since the rate of growth of labor inputs was overall slower than the rate of growth of capital inputs, total factor productivity in Latin American countries would be underestimated.

Total productivity growth was then recomputed using Bruton's data for the rates of growth of output and input, as shown in Table 2, Columns (1), (3) and (4), but employing as weights the mean labor and capital shares for the whole period considered, computed from our data. The series appear in Table 2. These results contrast sharply with those obtained by Bruton and validate our contention that he underestimated productivity in the Latin American countries examined. In all countries and all sub-periods, except in Chile for 1940–45, Bruton's estimates of total productivity were lower. In some instances Bruton's ratio of the rate of growth of total productivity to the rate of growth of output (r_a/r_p), Column (6), was as much as 50 percent lower than my estimates, Column (5).

III. *Conclusion*

This investigation demonstrated that Bruton's use of questionable factor shares led to an underestimate of the productivity change in four of five of the Latin American countries he investigated. The implications of this finding are not damaging for Bruton's contention that in the post-war period total productivity change in Latin American countries can be explained by changes in capacity utilization. Regressions run between the new productivity estimates and the rate of growth of output and inputs substantiate in broad terms Bruton's conclusions [2, p. 1109]. However, final judgment on this point must be reserved until detailed data on capacity utilization and on quality of input indices become available.

Both Bruton's and my estimates of total productivity change are obviously impure in the sense that they include other effects besides "... costless advances in applied

technology, managerial efficiency and industrial organization," which alone define productivity change [1, p. 764]. Jorgenson's and Griliches' recent study has demonstrated that, if proper account is taken of errors due to aggregation, and changes in capacity utilization and factor services are properly measured, the ratio of total productivity growth to the output growth of the U.S. private sector in the period 1945–65 was only .03 [6, p. 272]. This compares to .46, when these adjustments are not made, and to .60–.66 found by Bruton for a comparable period for the whole economy, but again without making proper adjustments [2, p. 1104].

Since the role of growth in total productivity in the United States and presumably in other advanced countries when measured properly is as small in absolute terms as shown by Jorgenson and Griliches, the fact that it might have been larger than in Latin American countries, even if correct, is of small consequence. It certainly does not appear to have the dire implications for Latin American growth that are suggested by Bruton.

This investigation has unveiled some interesting facts about factor shares in developing countries, a topic hitherto somewhat neglected. A considerable degree of stability in these shares was discovered and the functional distribution of income between labor and property was shown to be not greatly different from the distribution encountered in more advanced industrial countries.

References

1. M. Abramovitz, "Economic Growth in the United States," *Am. Econ. Rev.*, Sept. 1962, *52*, 762–82.
2. H. J. Bruton, "Productivity Growth in Latin America," *Am. Econ. Rev.*, Dec. 1967, *57*, 1099–1116.
3. E. F. Denison, "Income Types and the Size Distribution," *Am. Econ. Rev.*, May 1954, *44*, 254–69.
4. E. D. Domar, "On the Measurement of Technological Change," *Econ. Jour.*, Dec. 1961, *71*, 709–29.
5. D. G. Johnson, "The Functional Distribution of Income in the United States, 1850–1952," *Rev. Econ. Stat.*, May 1954, *36*, 175–82.
6. D. W. Jorgenson and Z. Griliches, "The Explanation of Productivity Change," *Rev. Econ. Stud.*, July 1967, *34*, 249–83.
7. C. Kerr, "Labor's Income Share and the Labor Movement," in G. W. Taylor and F. C. Pierson, eds., *New Concepts in Wage Determination*, New York 1957.
8. S. Kuznets, "Distribution of National Income by Factor Shares," *Economic Development and Cultural Change*, Apr. 1959, *7*, 1–100.
9. R. M. Solow, "A Skeptical Note on the Constancy of Relative Shares," *Am. Econ. Rev.*, Sept. 1958, *48*, 618–31.
10. Argentina, Consejo Nacional de Desarrollo, *Cuentas nacionales de la Republica Argentina*, Buenos Aires 1964.
11. Banco de la Republica Colombia, *Cuentas nacionales: 1950–1963*, Bogota 1963.
12. Chile, Corporacion de Fomento de la Produccion, *Cuentas nacionales de Chile: 1940–1954*, Santiago 1959.
13. United Nations, *Statistical Bulletin for Latin America*, Mar. 1965, New York 1965.

THE TWO GAP APPROACH TO AID AND DEVELOPMENT: COMMENT

By Henry J. Bruton*

The central idea of the "gap analysis" of Hollis Chenery and his various collaborators is that development tends to create situations

* The author is professor of economics, Williams College. He thanks Stephen R. Lewis, Jr. for useful comments on an earlier draft. He also acknowledges assistance provided under a research grant to Williams College by the Agency for International Development. The Agency is, of course, in no way responsible for any of the results reported here.

which, at various points in time, are characterized by a plentiful supply of all but one or a few of the factors "required" for continued development.[1] For these few, a gap between the quantity supplied and that required slows growth or halts it completely. When growth is thus limited by a bottleneck, there is underutilization of other factors. Foreign aid can then serve as a means of breaking the bottleneck, thereby permitting fuller utilization of all resources and a continuation of development.

Chenery *et al.*, concentrate most of their attention on a saving-investment gap and an import-export gap. A target rate of growth is postulated and a capital–output ratio is accepted as a datum. Hence a specific saving rate is derived as necessary to achieve the targeted growth rate. Similarly, a fixed relationship between imports and growth of output is postulated from which one may derive the level and rate of growth of imports required. A saving gap appears when the domestic saving rate is below the level necessary to permit the investment required to achieve the target, while imports are adequate. Aid covers the saving gap, and permits the achievement of the target. A trade gap appears however if with adequate savings, the flow of imports is below the required level. Here aid breaks the import bottleneck and permits the target to be reached. In this latter case, the key assumption is that the country is unable to transform its potential savings into exports. Furthermore in a typical developing country, one expects a specific time sequence of the bottlenecks.

In the exchange between John Fei and Gustav Ranis [8] on the one hand and Chenery and Alan Strout [7] on the other, primary attention was given to the meaning and empirical validity of the parameter values necessary to produce the time sequence of gaps specified by Chenery. The present note seeks to pose a prior question, namely the significance of the parameter values necessary to produce a distinction between the gaps at all. Essentially the point here is that aid is gap producing, not gap covering, and that accepting it as the latter can in fact impede, rather than facilitate, development. To demonstrate this, two issues are considered: (1) the existence of two distinct gaps during a given short interval of time, and (2) the origins of the separate gaps. These issues will be taken up in turn.

I. *The Existence of Two Gaps*

Chenery's major argument as to the existence of two gaps is familiar. Emphasis is given to a "structural" argument, i.e., in any given time period a developing economy can neither increase its exports nor decrease its imports without imposing underutilization on the economy. Export earnings for the bulk of products are largely determined by foreign demand conditions, and "a rapid increase in exports typically requires the development of new export products which is limited by productive capacity as well as organizational and institutional factors" [6, pp. 689–90]. Imports are required by the nature and limited flexibility of the productive system and of the composition of consumer demand. With such rigidity assumed, if the trade gap is larger than the saving gap, saving potential is "wasted" as resources released from consumer goods production can be used neither to produce capital goods nor exports. To change this structure requires extra resources (aid) and time.

1. The simplest assumption is that all (and only) capital goods are imported, that all exports are consumer goods, and that there are no changes in inventories of consumer goods or foreign exchange reserves. With this assumption, to try to save is to reduce consumption and actually to save is to do that and to export. Consider Figure 1. Let *dd* be the domestic demand curve for a typical export (consumer good) and *Cf* the foreign demand curve, i.e., foreign demand is perfectly elastic. The total demand curve (the curve relevant to the producers) is *daf*. The supply curve is *ss* and the quantity supplied at equilibrium price *OC* is *Cb* of which *Ca* is consumed domestically and *ab* is exported. The export of *ab* represents saving, the corresponding imports are capital, and the limitation on the rate of capital ac-

[1] The principal paper is [6], but there are a number of others equally useful. See especially [1], [5].

cumulation (and hence on growth in the Chenery model) is the saving rate. If the saving rate is increased, the domestic demand curve shifts leftward to $d'd'$, domestic consumption is reduced and exports and capital formation are increased by the same amount. Evidently there can be no trade gap distinct from the saving gap with these assumptions.

Suppose, however, that the foreign demand curve were vertical, as in Figure 2. The curve tt is obtained by summing horizontally dd and ff. In this event an increase in the saving rate, shown by the shift of dd to $d'd'$, has no effect on exports. The product price may tend to fall from OP_1, but this cannot help because the foreign demand curve is completely inelastic, and indeed the fall in price must induce an increase in the quantity demanded domestically until the new equilibrium is reached at price P_2, or unemployment will result. Evidently a similar result follows if ff has an elasticity greater than zero but less than unity over the relevant range. If the elasticity of ff exceeds unity and is less than infinite, and prices fall as domestic demand falls, the rate of capital formation will rise with an increase in the saving rate, but become increasingly expensive, i.e., a given increase in the rate of capital formation requires more saving (more resources devoted to export production) than at price P_1. At some point the curve will become inelastic and the rate of capital formation will have reached its maximum.

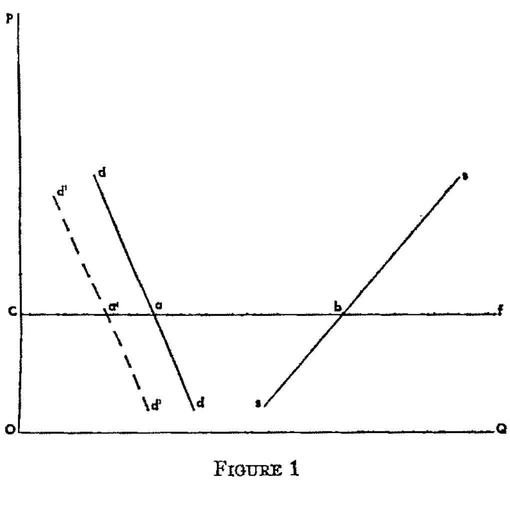


FIGURE 1

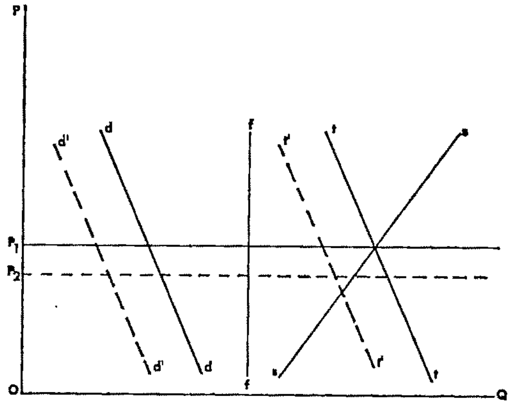


FIGURE 2

If the target rate of growth requires a higher growth rate of capital than this maximum, then a trade gap emerges which cannot be closed by increasing the saving rate.

This explanation of a two gap situation depends on assumptions about the elasticity of the foreign demand curve. There may be a problem on the supply side in a heavily protected domestic market. In Figure 3 the foreign demand curve is again horizontal at price OP_1. At this price, however, domestic suppliers are unwilling to export at all and survive only because they are protected from foreign competition. An increase in the saving rate will move the domestic demand curve to $d'd'$, but no increase in exports will be forthcoming. Devaluation will help in

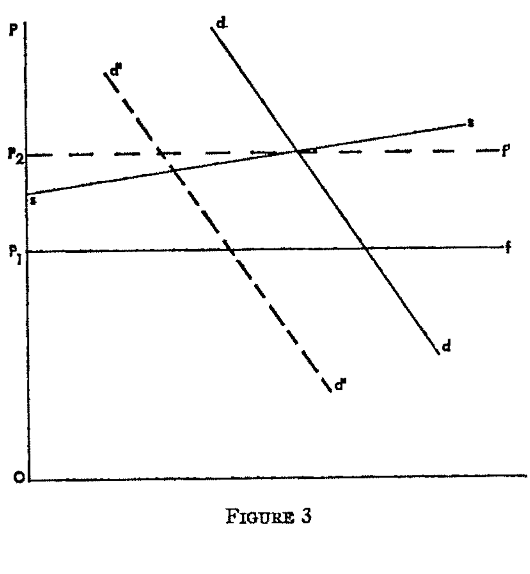


FIGURE 3

this case. In terms of the domestic currency, devaluation will result in the foreign demand curve rising, and a sufficient devaluation would then push P_1f to (say) P_1f', where increasing the saving rate will necessarily raise the rate of capital formation and the possibility of two gaps disappears.

A common situation in the developing world involves both Figures 2 and 3. Figure 2 applies to traditional exports, where foreign demand is quite inelastic and devaluation will reduce the rate of foreign exchange earnings. Such exports may constitute a large part of total exports. Figure 3 applies to a newly established nontraditional sector (probably, but not necessarily, manufacturing). It may face a horizontal (or almost horizontal) demand curve, and its costs may be such that it is priced out of the world market. This sector now constitutes a very small portion of total exports, but devaluation would raise its foreign exchange earnings. With a very large traditional sector and a small new sector, simple devaluation may not increase foreign exchange earnings, but devaluation plus a duty on traditional exports (or dual exchange rates) would. Without devaluation and the export duty there can indeed be two gaps, but they exist not for structural reasons, but because of the inability or unwillingness to pursue a policy that would eliminate the distinction between the gaps.

2. In the preceding argument the assumption was made that no capital goods sector existed in the developing country. Suppose now, however, that there is such a sector, but continue to suppose that only capital goods are imported. Obviously if a country has a capital goods sector, it may transform a part of its saving into capital without going through international trade. Hence, the rate of capital formation required to reach the targeted rate of growth of output can be achieved with a lower rate of exports than in the previous case. It may, however, still be "too" small, given the level of exports, and the foreign exchange bottleneck appears as a consequence of the demand situation described above.

Consider now a more subtle point. A capital goods sector is one in which machines, equipment, plant, and other forms of physical capital are produced, and as such it is reasonably well defined. So defined, its capacity to produce can clearly be quite small. Recent literature, however, has emphasized the importance to development of education, technical research, health, and a variety of factors other than physical capital. These sources of growth are not free, but use up investible resources, i.e., the use of resources to improve education, health, etc. is a form of capital formation. The "capital goods sector" must then include all activities which have the effect of increasing the productive capacity of the economy. With this broad definition of "capital goods sector," the notion of the fixity of size of the sector becomes virtually untenable. For one would then have to say that there is no use of investible resources that will raise the productive capacity of the economy.

Suppose foreign demand were such that there really was no way to increase foreign exchange earnings. Investible resources used to increase the capacity of the export sector will, in this case, have a social marginal product (*SMP*) of zero (or a capital-output ratio infinitely high). Resources used in other—domestic—capacity-increasing activities (improving or extending education, technical research, economic policy research) will have (in virtually all cases) a positive *SMP* (or a finite capital-output ratio). Evidently, then, investible resources should be applied to these latter activities. If, when available investible resources are applied in this way, the target rate of growth of *GNP* is not reached, the problem is that the saving rate is too low, i.e., there is again not two gaps, only one.

3. The assumption was made in both 1 and 2 that only capital goods were imported. Consider now the consequence of a given structure of production that requires imports of raw materials and spare parts to keep the existing capacity fully utilized.[2] For example, automobile assembly plants may

[2] One may also introduce consumer good imports but the consequence of reduced imports of these products on the argument is self-evident.

exist with no domestic source of steel, and to operate the assembly plants at any given level of capacity requires imports of steel at some specific rate. If steel imports are not available at this rate, the level of utilization of automobile assembly capacity must fall. To curtail imports for this purpose (maintenance imports) in order to step up capital goods imports means that the economy is penalized in the form of underutilization of existing plant and equipment. This situation then imposes a lower limit below which maintenance imports cannot be reduced without sacrificing the level of output as well as the rate of growth of output.

Even here, however, with a given structure of production, to assume that import requirements are independent of available policy measures is to give in much too easily. It is easy to find examples where domestic policies are such that a high rate of consumption of import intensive services occurs. This applies especially to the pricing of many public services, e.g., passenger transportation and electric power. The pricing of these (import intensive) services is such that their consumption is encouraged, where with more realistic pricing either greater savings or a shift in consumption would occur, either of which would contribute to resolving the trade gap. A less obvious point may be of greater relevance. If domestic demand for consumer products falls as policies to increase the saving rate are implemented, the appearance of excess capacity (at going prices) may, as noted, bring about reduced prices which might help a bit. The possibility that declines in domestic demand will lead entrepreneurs to make a greater effort to export is important in the present context. This greater effort takes the form of more intensive market efforts, attempts to maintain announced schedules, possibly financing arrangements, and the like. If domestic demand is strong, then producers have less incentive to search out export markets.[3]

II. *The Origins of the Two Gaps*

Although it is possible to devise a set of circumstances where saving cannot be transformed into capacity-creating activities, the possibilities are neither numerous nor empirically convincing. Indeed the arguments surely suggest quite strongly that, although the cost of growth may rise, the distinction between the two gaps as an empirical phenomenon is a rarity, and its explanation, where it does exist, as a structural phenomenon, i.e., a phenomenon not lending itself to short-term policy measures, is most unlikely. Emphasis may be placed on the fact that this conclusion does *not* depend on assumptions as to the empirical magnitudes of the several relevant elasticities. Neither does it depend on neoclassical assumptions as to flexibility and adaptability of all inputs. Such assumptions would of course eliminate the problem completely.

Suppose, however, that a structure does exist where it is meaningful to speak of two gaps; there is still a prior question: namely, where did this structure come from? If it emerged as a necessary and endemic consequence of the nature of the development process, that is one thing. If it emerged as a consequence of specific policy measures, that is, of course, another matter. Consider now the origins of a two gap situation.

Professor Chenery explains the origin of a two gap economy in terms of the effect of aid—supported exchange rates on investment allocation relative to that allocation called for after aid has been reduced or discontinued entirely. Chenery writes [6, p. 726],

> "If investment and other allocation decisions are based on the exchange rate that is appropriate for a substantial flow of aid, they are not likely to induce sufficient import substitution or increased exports to make possible a future reduction in the capital flow. Planning should be based on the higher equilibrium exchange rate that would be appropriate to a declining flow of aid in order for the necessary changes in the productive structure to be brought about in time."

After aid is reduced, the inflexibility and unadaptability of the system prohibits ad-

[3] Evidence on this point is, of course, not beyond question. Perhaps the most useful has to do with evidence that in periods of downturns or recessions, in some of the developing countries exports rise even though export prices do not fall or fall only moderately. This seems especially true for some of the Latin American countries. See [11] *passim*.

justments that would have to be made if the two gaps were not to appear. There are several reasons why this argument may be questioned.

The first point is purely empirical. Chenery cites Greece, Israel, Taiwan, and the Philippines as instances of successful uses of aid. There is no evidence, however, that in these countries investment allocation was based on the higher equilibrium exchange rate that is presumed to be appropriate after aid is reduced. In the absence of such evidence one may think that the success of these countries is due to factors other than the one isolated by Chenery.

More important than the empirical evidence, however, is the argument that the same exchange rate can be appropriate for a situation where aid is flowing in and for the situation after aid has been reduced. The central point seems *not* to be the exchange rate, but rather the composition of investment, especially the division between the consumer goods sector and the capital goods sector, and the rate of growth of productivity.

In the first part of this paper it was shown that, with a very narrow interpretation of capital and defining physical capital as the chief or sole source of growth, the size of the capital goods sector was strategic in the existence of a two gap situation. If the domestic capital goods sector were large enough, then the required savings could be, by definition, transformed into the physical capital necessary to produce the targeted rate of growth of output. Thus if the capital goods sector expands enough during the aid-receiving period, then in subsequent periods the two gap problem will not emerge.

Why doesn't the capital goods sector expand in the "balanced" way necessary to keep the foreign exchange gap from appearing? There are many possible reasons, but one is surely dominant and leads to a widely applicable conclusion. A developing country, seeking to industrialize, almost always proceeds by way of curtailing imports. In choosing the manufacturing sectors to protect, it is understandable that policymakers select sectors which are least disadvantaged in terms of costs or are least "essential" in some sense.[4] In a great variety of countries consumer durables fall in these categories and have become the major types of activities to replace imports. Rarely can one find a case where capital goods fall in the protected categories, as the difference between imported price and domestic costs is usually much larger than it is for consumer goods, and of course no one classifies capital goods as luxuries. So the economy has new consumer goods industries. It has used aid to import capital to build industries and possibly to import raw materials and spare parts to support them. The country has new activities the products of which it cannot export and which cannot be used to produce capital goods. As this process continues, the country acquires more activities of this sort. As the amount of investment rises, more imports of capital goods and more maintenance imports are demanded, and the kind of situation shown in Figures 2 and 3 emerges.

Suppose aid were maintained in sufficient quantities to support the investment for a number of years, and then reduced. Evidently, the investment rate would have to fall. Now does the gap problem emerge because of the exchange rate question referred to by Chenery? It is difficult to believe so. There is indeed a signal problem, but it is not a question of the exchange rate signal. It is rather the absence of a signal. Investing in the least disadvantaged sector creates activities whose relevance for a general equilibrium solution is purely coincidental, and one must fall back on a projection of the "structure" of the economy to ascertain the correct allocation of investment, i.e., the allocation that will prevent the growth process from grinding to a halt when aid is tapered off.

Suppose the activities which are initially least disadvantaged are all consumer goods industries. Then as aid falls and the bottleneck previously described emerges, this is a signal that investment should have occurred

[4] For a very lucid, thoroughly documented discussion of this point as it applies to Latin America, see Santiago Macario [9].

in the capital goods sector in order for the growth to continue. To produce capital goods domestically is by assumption, and in fact, more costly than producing consumer goods. If this were not the case, capital goods would have been produced initially domestically rather than consumer goods. As the country reaches the point where additional substitution of domestic production for imported consumer goods is not possible, it must begin to build capital goods. Then the cost of growth begins to rise. In this event, one would be saying not that the country had hit a foreign exchange barrier, but rather that the country was too poor to support its own development. This, however, is precisely the saving problem.[5] The only difference between this stage and the previous one is the size of the development bill, and the notion that the developing country can pay the lower price, but not the higher one. Obviously, this is a saving problem, not a trade gap problem.

If one argued that the country could not build capital goods because it did not have the labor skills, organization skills, and technical knowledge to do the job, this fact (assuming it to be a fact) is itself evidence of the wrong allocation of investment at some point in the past. Thus when Chenery writes that the "existing economic structure at any moment of time also limits the feasible growth of export earnings" [6, p. 689] and that "a rapid increase in exports typically requires the development of new export products, which is limited by productive capacity as well as organizational and institutional factors" [6, pp. 689–90] and that these difficulties can be removed only over time, he is saying that investment allocation in the past has now proved to be wrong. It was wrong in the sense that it produced a "structure" that cannot now be fully exploited. Where the difficulties are now skills and technology, the conclusion is that earlier investments should have been in activities that developed these aspects of the economy.

What emerges from these arguments is quite orthodox, and very much in line with what Chenery has taught us in his several papers on investment allocation criteria [3] [4]. The point here is to emphasize that the special difficulties attributed by Chenery to exchange rates with aid and after the end of aid are not really the heart of the issue. The problem is simply the old one of achieving the optimal saving rate and an investment composition that follows the social marginal productivity criterion. This criterion, broadly defined,[6] is of course difficult to apply in any circumstance, but that is not the point. In this case one must project the structure and allocate the investment to seek to prevent gaps in this structure from emerging. The allocation criterion must be gap prevention rather than least disadvantaged or some other cost based argument.

There is a final point of great relevance. Suppose a trade gap appears due either to distortions emphasized here or to Chenery's structural arguments. Aid can then provide the resources to correct or change this set of circumstances. Aid can also do something else. It can provide the resources with which an economy can continue to function acceptably *without* bringing about the elimination of the distortions or changing the structure. By relieving the pressure on the system, aid may also reduce not only the incentive to make painful changes, it may hide the location of the right allocations. The point here is, not that international aid should be reduced, but rather that its effective use places great demands on the policy-makers as to the understanding of the development process. Policies built around the assumption that a trade gap is a necessary condition of development may impede that understanding.

[5] The usual view of the development process in the USSR illustrates this point. The Soviet Union depended little, if at all, on imports, and did not hit a trade barrier. In the simplest terms, the explanation is that the Soviet leaders were willing and able to force a level of saving on the economy that did permit the construction of a large and costly capital goods sector. More directly, the great capacity of the Soviet government to force saving solved the trade gap problem. On the general argument see Winston [10].

[6] "Broadly defined" means inclusive of the "side" or "indirect" effects, e.g., effect on saving rates, on labor training, population growth, now commonly included in allocation analysis. See [2, Chapter 15].

III. *Conclusion*

This paper has argued three things: (1) the distinction between the two gaps is due to particular policies that themselves are growth impeding, and not to some inherent characteristic of the development process; (2) that foreign aid based on the assumption that there are two gaps may tend to perpetuate these growth impeding policies; and (3) that the traditional investment allocation issues remain strategic in development policymaking.

REFERENCES

1. I. ADELMAN AND H. CHENERY, "Foreign Aid and Economic Development: The Case of Greece," *Rev. Econ. Stat.*, Feb. 1966, *48*, 1–19.
2. H. BRUTON, *Principles of Development Economics*, Englewood Cliffs, 1965.
3. H. CHENERY, "The Application of Investment Criteria," *Quart. Jour. Econ.*, Feb. 1953, *67*, 76–96.
4. ———, "Comparative Advantage and Development Policy," *Am. Econ. Rev.*, Mar. 1961, *51*, 18–51.
5. ———, AND M. BRUNO, "Development Alternatives in an Open Economy: The Case of Israel," *Econ. Jour.*, Mar. 1962, *72*, 79–103.
6. ———, AND A. STROUT, "Foreign Assistance and Economic Development," *Amer. Econ. Rev.*, Sept. 1966, *56*, 679–732.
7. ———, "Reply," *Am. Econ. Rev.*, Sept. 1968, *58*.
8. J. C. H. FEI AND G. RANIS, "Foreign Assistance and Economic Development Revisited," *Am. Econ. Rev.*, Sept. 1968, *58*.
9. S. MACARIO, "Protectionism and Industrialization in Latin America," *Econ. Bull. for Latin America*, Mar. 1964, *9*, 61–102.
10. G. C. WINSTON, "Consumer Goods or Capital Goods—Supply Consistency in Development Planning," *Pakistan Develop. Rev.*, Autumn 1967, *7*, 348–378.
11. ECONOMIC COMMISSION FOR LATIN AMERICA, *The Process of Industrial Development in Latin America* (Mimeo.), Santiago, 1965.

THE TWO GAP APPROACH TO AID AND DEVELOPMENT: A REPLY TO BRUTON

By HOLLIS B. CHENERY*

I find myself in agreement with most of H. J. Bruton's analysis but in disagreement with his main conclusions. The source of our disagreement lies in Bruton's reliance on the assumptions of neoclassical equilibrium, whereas the two gap approach assumes that underdeveloped economies are considerably less flexible. I will try to identify the main differences in assumptions before taking up their policy implications.

I

"Gap analysis" is a by-product of attempts to determine the actual policy alternatives facing underdeveloped countries.[1] It focuses on the probable limits to accelerated growth and the extent to which they can be overcome through the use of external resources. In this framework, the two gap diagnosis has two different meanings: (i) as a description of an existing condition of structural disequilibrium; (ii) as a potential limit to future growth. Although the widespread diagnosis of existing disequilibrium provides evidence that the problem is real and important, it is the distinction between the trade and savings limits in the future that is more relevant for policy.

It has become customary to accept a government's limited ability to increase taxes and otherwise affect savings as given, but to regard its limited success in increasing exports or curtailing imports as merely evidence of bad policy. There seems to be no a

* Center for International Affairs, Harvard University.

[1] Most of the applications deal with individual countries [1], [3], [4]. My paper with Strout [5] omitted this detailed discussion in trying to draw general conclusions.

priori reason to make such a strong distinction, however. In both cases, the limits to government action are set by its diagnosis of the problem, the likely response of the economy, and the political acceptability of the results. If one is planning for accelerated development over the next five or ten years, the question becomes an empirical one: how rapidly can both the trade and savings limits be raised with optimal policies, including the use of external resources?[2]

Equilibrium theory has little to say about such questions. Two gap analysis assumes that it will be difficult or impossible to maintain an optimal relation between the two limits, particularly if we start from a situation in which they are not equally constraining, and therefore incorporates them explicitly in the analysis of capital inflows.

A diagnosis of past causes of either balance of payments disequilibrium or low savings rates is important primarily for the guidance that it provides as to future possibilities for improved performance. Whether or not disequilibrium in the payments balance could have been avoided by having had a higher peso-dollar exchange rate for the past ten years does not determine whether the current problem in country X is structural. The effects of an overvalued rate or other misguided policies are incorporated in the existing distorted structure of production and trade. The fact that it requires a reallocation of investment and other changes extending beyond the short run to expand the trade limit makes the problem structural, whatever its origin.

In summary, two gap analysis assumes that past policies may have been imperfect and that it will take time to bring about the increase in savings and exports needed to support accelerated growth. It also incorporates estimates of the effects over time of policies designed to accomplish these structural changes. Bruton argues that the problem would not exist if countries had followed ideal policies in the past. Whether or not this is true, it is not particularly relevant to a description of actual policy alternatives.[3]

II

Bruton's main contribution to the two gap discussion is his suggestion that there are forms of productive investment in human resources, research, etc., that do not require imported capital goods. So long as this type of investment can be expanded with positive productivity, the trade limit cannot prevent the translation of potential savings into investment and the possibility of excess savings disappears. However, the effect of the trade limitation is still reflected in the difference between the lower growth resulting from this constraint on the composition of investment and the higher growth produced with an optimal composition. The magnitude of this reduction is shown below to be the relevant empirical question, not whether the marginal product of investment in the trade-limited case is zero or "very low."

Several of the two gap papers [2], [4], [5] have considered ways in which an increase in output might be achieved by fuller use of potential savings. They conclude that additional investment in exports or import substitutes with a lower return to capital is the most likely possibility, and subsequent analysis was based on that assumption. Bruton's suggestion provides an indirect way of substituting domestic savings for imported capital goods by changing the composition of investment. We all conclude that it will be efficient under most assumptions to eliminate the "... gap between the gaps ..." [5, p. 701]. However, using all the potential

[2] In the first two gap paper—on Israel [3]—the question was reversed: how rapidly could the large inflow of capital be reduced while maintaining the high rate of growth. The evidence that for several years to come the trade gap would be more of a limiting factor than the savings gap was quite persuasive once the implied shifts in investment allocation and the required growth of exports were examined. In this case, the demonstration did not rely on a finding of a difference between the two gaps at the existing level of capital inflow.

[3] The discussion is reminiscent of the debate over the "necessity" of Keynesian unemployment in an earlier day.

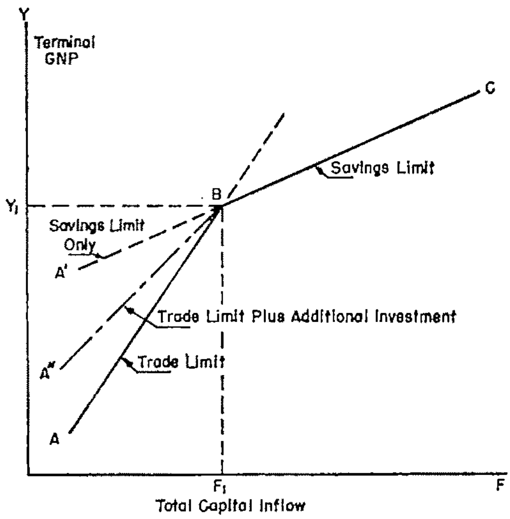


FIGURE 1
PRODUCTIVITY OF EXTERNAL CAPITAL

savings is not equivalent to getting rid of the trade limit nor does it reduce the desirability of external capital.

III

The question remains as to whether Bruton's results affect the other implications of the two gap formulation as to the productivity of external assistance and the allocation of investment. The effect of Bruton's "investment in education" on the two gap description of development alternatives is shown in Figure 1, which is similar to the empirically based aid productivity functions in [2], [3], [4]. If there were no trade limit, the contribution of external capital to growth of GNP would be given by the curve $A'BC$. If a trade limit is assumed with no use for surplus savings, the productivity of external capital is lowered to ABC and the aid requirement for a target GNP below Y_1 is increased. Either Bruton's additional investment in education or my trade-improving investment results in an intermediate segment $A''B$, whose distance below $A'B$ depends on the fall in the marginal productivity of the additional investment.

Several two gap studies concluded that it would normally be efficient to borrow at least up to point B, where the marginal productivity of external capital drops. Substituting "low productivity" for "zero productivity" uses of savings at the margin is not likely to change this conclusion, since the productivity of aid up to an amount F_1 will be significantly greater below that point on both Bruton's and my assumptions.[4]

Bruton and I agree that the trade limit typically arises when easy (and cheap) import substitution possibilities have been exhausted and traditional exports no longer suffice to finance the remaining import needs. Although external capital may provide a short-term solution, investment must be redirected to either new (often manufactured) exports or to new import substitutes, either in capital goods or intermediate products. Bruton's treatment of these alternatives is curiously one sided. Although in Section I he points out the possibility of expanding manufactured exports if the exchange rate is devalued, the only solution suggested in Section II is to substitute for imports of capital goods. I take it that we would agree on how this choice should be made, either by linear programming in a general equilibrium framework (as in [2]) or by a partial approximation that compares the cost of dollar earning to dollar saving. In either case, the shadow price of foreign exchange is a key to the right choice. Empirically, I would argue that it will usually point to increased exports rather than to capital goods as the cheaper way to expand the trade limit. Hence my emphasis on this variable as a guide to future equilibration of the two gaps.[5]

[4] In the linear programming formulation of [4], the value of external capital is a weighted sum of the shadow prices of savings and foreign exchange. Along the segment AB the shadow price of savings is zero, while along $A''B$ both prices are positive. The importance of the trade limit is determined by the relative weight of exchange saving in determining the value of external capital.

[5] For some reason Bruton has taken my discussion of the role of the exchange rate in guiding resource allocation in the future to be my explanation of the causes of existing two gap disequilibria. My intention was to argue that the equilibrium exchange rate is a function of

Although two gap disequilibrium seems in theory to be no more inevitable than Keynesian unemployment, it remains to be seen whether economists will have greater success in helping governments to avoid it. Up to now the record is not impressive. My own guess is that the exchange rate alone will prove as inadequate as the interest rate has been in the Keynesian situation, and that more comprehensive policy packages will be needed to secure the needed redirection of resources. Until this triumph of economic science takes place, there will be some use for two gap thinking.

---

the rate of capital inflow—as is amply demonstrated in [2]—and that its rise with declining capital inflow provides a measure of both the increasing cost of development and a guide to the allocation of investment between export expansion and import substitution (including the capital goods industry). Since Bruton quotes my other writings on investment criteria with approval, I can only conclude that the statement in [5, p. 726] is obscure.

References

1. I. Adelman and H. Chenery, "Foreign Aid and Economic Development: The Case of Greece," *Rev. Econ. Stat.*, Feb. 1966, *48*, 1–19.
2. M. Bruno, "A Programming Model for Israel," in Adelman and Thorbecke (eds.), *The Theory and Design of Development*, Baltimore 1966.
3. H. Chenery and M. Bruno, "Development Alternatives in a Open Economy: The Case of Israel," *Econ. Jour.*, Mar. 1962, *72*, 79–103.
4. ——— and A. MacEwan, "Optimal Patterns of Growth and Aid: The Case of Pakistan," in Adelman and Thorbecke (eds.) *The Theory and Design of Development*, Baltimore 1966.
5. ——— and A. Strout, "Foreign Assistance and Economic Development," *Amer. Econ. Rev.*, Sept. 1966, *56*, 679–732.
6. ——— and———, "Reply," *Amer. Econ. Rev.*, *58*, Sept. 1968.

NOTES

An Invitation from Kenneth E. Boulding to Join a New Association for the Study of the Grants Economy (ASGE)

The grants economy is becoming a sizeable part of all modern economic systems and is an essential instrument in the study of the integrative system. Its theory is an essential complement to the theory of exchange and is in no way inconsistent with it. It does involve expanding the underlying psychology of economics beyond such concepts as the Paretian optimum, to include concepts of interrelatedness of utility functions through benevolence or malevolence and to include certain elements of what is being called the "integrative system," such as the status and identity of individuals and the communities with which they identify. The international and comparative study of the grants economy is an important element of the problem.

The membership will be both international and interdisciplinary and open to all who are interested. The subscription fee of $2 may be sent, or membership forms may be obtained by contacting the Secretary-Treasurer, Professor Martin Pfaff, Guest Scholar, The Brookings Institution, 1775 Massachusetts Ave., N.W., Washington, D.C. 20036.

ASGE will hold a symposium in Boston, Massachusetts December 26–27, 1969 as part of the annual meeting of the American Association for the Advancement of Science, and coordinated by its section of Social and Economic Sciences.

The interests of the syposium include the various areas outlined above, and related topics. Thus, sessions are planned in order to study the theoretical aspects, especially measuring techniques, of unilateral transfers; and offer empirical examinations regarding motivations and attainments of various types of unilateral transfers.

Economists and other social scientists who would like to contribute papers are invited to forward titles and abstracts of 300 words to Professor Janos Horvath, Program Coordinator, Department of Economics, Butler University, Indianapolis, Indiana 46208, not later than August 25.

Travel grants to residents of Australia, New Zealand, Singapore, South Africa and the West Indies were announced by Carnegie Corporation of New York.

Z. S. Gurzynski, senior lecturer in economics, University of Cape Town, to study teaching methods and research in urban and regional economics in the U.S. and Canada. Arriving December 1969.

Marcell Kooy, senior lecturer in economics, University of Cape Town, to study teaching of economic history and student-faculty relations in the U.S. Arriving September 1970.

Announcement of foreign economists available for appointments in U.S. universities and colleges under provisions of the Fulbright-Hays Act for the academic year 1969–70.

Yugoslavia: Ljubivoj Blagojevic, Eftim Bojadzievski, Juraj Padjen, Mladen Stretenovic, Vinko Trcek, and Japan: Makoto Sakurabayashi.

Inquiries should be addressed to: Miss Grace E. L. Haskins, Program Officer, Committee on International Exchange of Persons, Conference Board of Associated Research Councils, 2101 Constitution Avenue, N.W., Washington, D.C. 20418 or via telephone—Area Code 202, 961-1648.

The economics department in the School of Commerce and the economics department in Washington Square College have merged into a single undergraduate economics department at the Washington Square campus of New York University.

The 54th annual meeting of the Association for the Study of Negro Life and History will be held Thursday–Sunday, October 9–12, 1969 at the Tutwiler Hotel, Birmingham, Alabama.

Persons interested in proposing sessions or papers should write to the Program Chairman, Walter Fisher, Department of History, Morgan State College, Baltimore, Maryland 21212.

The 20th Alaska Science Conference will be held August 24–27, 1969 at the University of Alaska campus, near Fairbanks, under the auspices of the Alaska Division, American Association for the Advancement of Science. Conference symposia and panels will be organized around the theme "Change in the North: People Resources, and Environment."

For information about submission of papers or attendance, contact Victor Fischer, Conference Chairman, University of Alaska, College, Alaska 99701.

The National Institute of Social and Behavioral Science will hold sessions for contributed papers at the 136th annual meeting of the American Association for the Advancement of Science in Boston December 26–31, 1969. Economists interested in presenting a paper at these sessions are invited to forward titles and abstracts of some 300 words to Donald P. Ray, Director, National Institute of Social and Behavioral Science, 863 Benjamin Franklin Station, Washington, D.C. 20044 by August 25.

Frank W. Taussig Award Winner

The first winner of the Frank W. Taussig award, sponsored by Omicron Delta Epsilon, is Mr. David S. Davenport of Amherst College. His winning essay was

"Collusive Competition in Major League Baseball: Its Theory and Institutional Development."

Mr. Davenport's paper was selected for the award by a board consisting of Professors, Kenneth J. Arrow, Kenneth E. Boulding, Milton Friedman, Paul A. Samuelson, and Egon Neuberger (Editor).

The Center for Law and Behavioral Science at the University of Wisconsin announces the 1969 Summer Institute in Behavioral Science and Law.

Funded by the National Science Foundation, the Institute will provide graduate students and law students with an opportunity to explore the problems of applying behavioral science to the study of law in society.

The curriculum for the Institute will include three basic graduate courses: "Law and Social Science", "Problems in Deviance and Criminal Responsibility", and "Basic Methods of Empirical Inquiry".

Students who are admitted will receive a stipend and travel allowance, and may receive graduate credit for courses taken. For further information, contact Professor Stuart A. Scheingold, Director, Summer Institute in Behavioral Science and Law, Department of Political Science, 406 North Hall, The University of Wisconsin, Madison, Wisconsin 53706.

Note on Exchange Advertisements

At the March 1969 meeting of the Executive Committee of the American Economic Association, it was decided to discontinue exchange advertisements with other economic journals. These ads are no longer needed because of the complete service provided by the Journal of Economic Literature, which prints titles and abstracts of articles in the major economic journals.

Deaths

Harold Abel, Denver, Colorado, February 16, 1969.

Adolph G. Abramson, Jenkintown, Pennsylvania.

Alexander Brody, assistant professor of economics, The City College of the City University of New York, October 20, 1968.

James E. Chace, Jr., Gainesville, Florida, December 12, 1968.

Dorothy Wolff Douglas, visiting professor of economics, Hofstra University, December 9, 1968.

John T. Harris, Birmingham, Alabama, December 18, 1968.

Karl Karsten, Washington, D.C., May 23, 1968.

Karl E. Lachmann, New Gardens, New York, August 30, 1968.

Charles Lemelin, Quebec, Canada, September, 1968.

Gabriel Lundy, head, agricultural economics department, South Dakota State University, 1936–1949, November 2, 1968.

Kuldip S. Mali, professor of economics, University of Omaha, July 1, 1968.

F. Eugene Melder, professor of economics, Clark University, December 27, 1968.

Jacob Perlman, Washington, D.C., April 8, 1968.

Michael A. Plesher, Phoenix, Arizona, September 12, 1968.

David J. Saposs, Washington, D.C., November 13, 1968.

Arthur Schatzow, Silver Spring, Maryland, October 1, 1968.

Andrew B. Schmidt, professor emeritus of economics, University of Arizona, February 22, 1969.

Julius T. Wendzel, Richland, Michigan, November 27, 1968.

Mary C. Wing, Washington, D.C., September 28, 1968.

Retirements

Robert J. Barr, professor emeritus, Marquette University.

Roderick B. Crane, professor of economics, University of Nebraska, Omaha, June 1969.

Walter Froehlich, professor emeritus, Marquette University.

Max Gideonse, professor of economics, Rutgers-The State University, June 1969.

William S. Hopkins, professor of economics, University of Washington, Seattle, June 1969.

Mortin J. Plotnik, professor of economics, Illinois Wesleyan University, June 1969.

Geoffrey S. Shepherd, professor of economics, Iowa State University, May 1969.

Verne F. Simons, associate professor of accounting, Rice University, June 1969.

Hans Singer, United Nations: chair, Institute of Development Studies, University of Sussex, England.

Harold F. Williamson, professor of economics, Northwestern University, June 1969.

Visiting Foreign Scholars

Josè Encarnaciòn, Jr., University of the Philippines: visiting professor of economics, University of Wisconsin-Madison.

Max Hartwell, Nuffield College, Oxford: visiting professor, University of Washington, Seattle.

Fred C. Hung, University of Hawaii, Honolulu: visiting associate professor, University of Washington, Seattle.

Eric L. Jones, visiting professor of economics, University of Reading, England: Northwestern University.

Meheroo F. Jussawalla, Hyderabad, India: visiting lecturer in economics, Hood College, Spring 1970.

Michael Landsberger, Jerusalem, Israel: visiting lecturer of economics, University of Pennsylvania, fall 1969.

John E. La Tourette, SUNY Binghamton: visiting scholar of economics, Arizona State University, spring 1969.

Richard G. Lipsey, University of Essex: visiting professor of economics, University of British Columbia, fall 1969.

Lachlan McGregor, Monash University: visiting associate professor of economics, Northwestern University.

Edward Nell, University of East Anglia, Norwich,

England: advanced theory, New School for Social Research.

Ivor Pearce, University of Southampton: visiting professor of economics, University of Pennsylvania, fall 1969.

John D. Pitchford, Australian National University: visiting professor, department of economics and commerce, Simon Fraser University.

Johannes Antonious Ponsioen, Institute of Social Studies, the Hague: visiting professor socioeconomic development, University of New Hampshire, fall 1968.

James Richmond, University of Essex: visiting assistant professor of economics, Northwestern University.

Kazuo Sato, United Nations: visiting professor of economics, Massachusetts Institute of Technology.

N. Eugene Savin, University of Essex: visiting assistant professor, Northwestern University.

Erich Schneider, Kiel University: visiting professor of economics, University of Pennsylvania, fall 1969.

Keith Sloane, Australian National University: visiting lecturer of economics, University of Pennsylvania, fall 1969.

C. Christian von Weizsäcker, Heidelberg University: visiting professor of economics, Massachusetts Institute of Technology.

David Wall, University of Sussex: visiting research specialist, language and area center for Latin America, University of Wisconsin-Milwaukee, summer 1969; visiting fellow in economics, University of Chicago.

Promotions

Robert T. Averitt: professor, Smith College.

Burley V. Bechdolt, Jr.: assistant professor, California State College at Los Angeles.

Jere R. Behrman: associate professor of economics, University of Pennsylvania.

David A. Belsley: associate professor of economics, Boston College.

Donald B. Billings: assistant professor of economics, California State College, Long Beach.

Thomas Bruce Birkenhead: associate professor of economics, Brooklyn College.

Edwin Burmeister: associate professor of economics, University of Pennsylvania.

Leonard Cain: associate professor of economics, The Catholic University of America.

John K. Chang: associate professor of economics, Lafayette College.

Curtis A. Cramer: associate professor, The University of Wyoming.

Karen Davis: assistant professor of economics, Rice University.

Eugene J. Devine: assistant professor of economics, Simon Fraser University.

Gordon K. Douglass: professor of economics, Pomona College and Claremont Graduate School.

Michael K. Evans: associate professor of economics, University of Pennsylvania.

Mark Z. Fabrycy: associate professor of economics, New York University.

Duncan K. Foley: associate professor of economics, Massachusetts Institute of Technology.

Stanley Friedlander: associate professor, The City College of the City University of New York.

Jean Wilburn Gooch: associate professor of economics, Barnard College, Columbia University.

Herschel I. Grossman: associate professor of economics, Brown University.

Peter M. Gutmann: professor of economics, Bernard M. Baruch College of the City University of New York.

Oswald Honkalehto: associate professor of economics, Colgate University.

Frank Hsiao (Sheng-tieh): associate professor, University of Colorado.

Erich Isaac: professor, The City College of the City University of New York.

George Jensen: associate professor of economics, California State College, Los Angeles.

Robert Jones: assistant professor, The City College of the City University of New York.

F. Thomas Juster: vice president-research, National Bureau of Economic Research.

J. William Leasure: professor of economics, San Diego State College.

Mildred G. Massey: professor of economics, California State College, Los Angeles.

Laurence J. Mauer: associate professor of economics, Northern Illinois University.

Abraham Melezin: associate professor, The City College of the City University of New York.

Constantine Michalopoulos: associate professor of economics, Clark University.

Se-Hark Park: associate professor, Marquette University.

Harold Petersen: associate professor of economics, Boston College.

Robert A. Pollak: associate professor of economics, University of Pennsylvania.

Ibrahim I. Poroy: associate professor of economics, San Diego State College.

Richard L. Porter: professor of economics, Oklahoma State University.

Hyman Sardy: associate professor of economics, Brooklyn College.

Samuel H. Talley: associate professor of economics, University of Maine.

John A. Tomaske: associate professor of economics, California State College, Los Angeles.

Theodore Walther: associate professor of economics, Bates College.

Oliver E. Williamson: professor of economics, University of Pennsylvania.

Alexander Woroniak: associate professor of economics, The Catholic University of America.

Kozo Yamamura: professor of economics, Boston College.

Administrative Appointments

Pekka Ahtiala: dean, faculty of economics and administration, Tampere University.

Edward Ames: chairman, department of economics, State University of New York, Stony Brook.

Gordon C. Bjork: president, Linfield College.

Forrest C. Blodgett: assistant business manager, Linfield College.

Levi Carlile: associate professor and chairman of the economics department, Linfield College.

Jerome B. Cohen: dean of the School of Business Administration, Bernard M. Baruch College of the City University of New York.

Harry L. Cook: chairman, department of economics, Southern Oregon College.

John S. Day: dean, Purdue University School of Business.

Daniel E. Diamond: vice dean, School of Commerce, New York University.

Robert L. Heilbroner: chairman, department of economics, New School for Social Research.

T. Edward Hollander: vice chancellor for budget and planning, The City University of New York.

Pamela Haddy Kacser, American University: chief, division of economic analysis, Office of the Secretary, Department of Health, Education and Welfare.

Rene P. Manes: associate dean, Purdue University School of Business.

Francis B. McCormick: associate chairman of the department of agricultural economics and rural sociology, Ohio State University, December 1968.

Noah M. Meltz, University of Toronto: acting chairman of the social sciences, Scarborough College.

Sid Mittra: acting chairman, economics department, Oakland University.

Donald A. Moore: chairman of the department of economics, California State College, Los Angeles.

Max Myers: director, Institute of Social Sciences for Rural-Urban Research and Planning, South Dakota State University.

James R. Nelson: chairman, department of economics, Amherst College.

Herbert E. Newman: chairman, department of economics and sociology, Hood College.

Jan Parker: chairman, department of economics, Sweet Briar College.

Alberto M. Piedra: chairman, department of economics, The Catholic University of America.

Bruce N. Robinson: assistant to the president, Haverford College.

Richard L. Ruth, Marquette University: associate professor, chairman of department of economics, St. Mary's University, Halifax.

Louis H. Schuster: dean of faculty, Mattatuck Community College.

Jack W. Skeels: associate dean and professor of economics, College of Liberal Arts & Sciences, Northern Illinois University.

Lloyd Swenson: director of financial aid, Linfield College.

John E. Thompson: head of the economics department, South Dakota State University.

Gene B. Tipton: associate chairman, department of economics, California State College, Los Angeles.

Leland B. Yeager: acting chairman, department of economics, University of Virginia.

John H. Young: dean, faculty of arts, University of British Columbia.

Elliot Zupnick: associate dean, Graduate Center, City University of New York.

Appointments

Frank J. Alessio: assistant professor of economics, University of Arizona.

Clark Lee Allen, Southern Illinois University: professor, Florida Presbyterian College.

James Anderson: instructor, Boston College.

Leslie Aspin: assistant professor of economics, Marquette University, and consultant to the Office of the Joint Economic Committee.

Raymond J. Ball: assistant professor of accounting, Graduate School of Business, University of Chicago.

Anton P. Barten: visiting professor of econometrics, Graduate School of Business, University of Chicago.

Richard N. Bean: assistant professor of economics, University of Houston.

Charles A. Bennett: instructor of economics, Gannon College.

Trent Bertrand, Cornell University: assistant professor, Johns Hopkins University.

Michael H. Best, University of Oregon: assistant professor, University of Massachusetts.

Robert C. Blattberg: assistant professor of business, economics and marketing, Graduate School of Business, University of Chicago.

William J. Breen: associate professor of economics, Purdue University.

Ronald B. Brooks: assistant professor of mathematical economics, Graduate School of Business, University of Chicago.

Douglas M. Brown: assistant professor of economics, Northeastern University.

F. Lee Brown: assistant professor of economics, University of New Mexico.

Kenneth M. Brown: assistant professor of economics, University of Notre Dame.

John A. Carlson: professor of economics, Purdue University.

Steven N. S. Cheung, University of Chicago: associate professor of economics, University of Washington, Seattle.

Peter Clark: instructor, Boston College.

Bruce C. Cohen: associate professor of economics, Northeastern University.

Jerome Phillip Cooper: assistant professor of business, economics and finance, Graduate School of Business, University of Chicago.

Thomas Cowing: lecturer in economics, State University of New York–Binghamton.

George D. Craig, Louisiana State University: University of Illinois.

William S. Dawes: assistant professor of economics, State University of New York, Stony Brook.

B. Curtis Eaton: assistant professor of economics, University of British Columbia.

Robert F. Engle: assistant professor of economics, Massachusetts Institute of Technology.

Robert G. Evans: assistant professor of economics, University of British Columbia.

George C. Eads, Harvard University: assistant professor, Princeton University.

Isaac Ehrlich: assistant professor in business eco-

nomics, Graduate School of Business, University of Chicago.

Hillel J. Einhorn: assistant professor in behavioral sciences, Graduate School of Business, University of Chicago.

Wilfred J. Ethier, University of Rochester: assistant professor of economics, University of Pennsylvania.

Robert W. Fenton: research fellow, department of economics and commerce, Simon Fraser University.

Robert J. Flanagan: assistant professor of labor economics, Graduate School of Business, University of Chicago.

Denise Rosemary Ford: assistant professor of economics, University of British Columbia.

Derek James Ford: assistant professor of economics, University of British Columbia.

Graeme H. Forrester: lecturer, department of economics and commerce, Simon Fraser University.

Donald R. Fraser, Federal Reserve Bank of Dallas: assistant professor of economics and finance, University of Texas, El Paso.

Raymond O. Gaarder: livestock marketing specialist, economics department, South Dakota State University.

Carl M. Gambs, Yale University: instructor in economics, Michigan State University.

Neil Garston, Brown University: instructor of economics, Trinity College.

Gilbert R. Ghez, National Bureau of Economic Research: University of Chicago.

Nicholas J. Gonedes: assistant professor of accounting, Graduate School of Business, University of Chicago.

Kenneth Gordon, Northwestern University: assistant professor, University of Massachusetts.

Nancy Gordon, Stanford University: assistant professor of economics, Graduate School of Industrial Administration, Carnegie-Mellon University.

Paul R. Gregory, Harvard University: assistant professor of economics, University of Oklahoma.

Ronald E. Grieson: assistant professor of economics, Massachusetts Institute of Technology.

Michael Grossman: research associate, Graduate School of Business, University of Chicago.

Theodore Groves, Jr., University of California, Berkeley: assistant professor, University of Wisconsin, Madison.

Samuel Gubins: assistant professor of economics, Haverford College.

Alan F. Gummerson: assistant professor of economics, Clark University.

Michael G. Hadjimichalakis, University of Rochester: assistant professor of economics, University of Washington, Seattle.

Walter W. Haessel: research associate, Iowa State University.

Thomas W. Hall: assistant professor of economics, Bowling Green State University.

Daniel S. Hamermesh, Yale University: assistant professor, Princeton University.

David A. Hansen: acting assistant professor of economics, Linfield College.

Niles Hansen: professor of economics, and director of Center for Economic Development, The University of Texas, Austin.

Charles K. Harley: assistant professor of economics, University of British Columbia.

Gerald A. Harrison: research associate, department of economics, Iowa State University.

Kichiro Hayashi: instructor in economics, DePauw University.

Walter P. Heller, Stanford University: assistant professor of economics, University of Pennsylvania.

George Henry: associate professor of economics, Colgate University.

Donald R. Herzog: professor of production management and marketing, coordinator of graduate programs in business, Chico State College.

John P. Herzog, Claremont Graduate School: professor, department of economics and commerce, Simon Fraser University.

Thomas J. Hogarty: assistant professor of economics, Northern Illinois University.

Ralph W. Huenemann: assistant professor, department of economics, University of British Columbia.

A. M. Huq: professor of economics, University of Maine.

John P. Huttman, Sonoma State College: assistant professor, department of economics and commerce, Simon Fraser University.

Eugene L. Jaffe: economist, U. S. Marine Corps.

David E. Kidder: assistant professor of economics, Northeastern University.

Alan Kirman, Princeton University: assistant professor, Johns Hopkins University.

Roland H. Koller, II: assistant professor of economics, Brigham Young University.

Stanley E. Kowalski: assistant professor of economics, Old Dominion College.

Walter Krause, professor of economics, The University of Iowa: appointed John F. Murray professor of economics.

Mark L. Ladenson, Northwestern University: Instructor in economics, Michigan State University.

James W. Land, St. Thomas University: associate professor of economics, Rice University.

W. Cris Lewis, Iowa State College: assistant professor of economics, University of Oklahoma.

Chong K. Liew, University of California, Berkeley: assistant professor of economics and economist for Bureau for Business and Economic Research, University of Oklahoma.

James D. Likens, University of Minnesota: assistant professor of economics, Pomona College.

Louis Maccini, Northwestern University: assistant professor, Johns Hopkins University.

Richard B. Mancke: assistant professor of business economics, Graduate School of Business, University of Chicago.

Julius Margolis, Stanford University: professor of Public Policy Analysis and Director of Fels Institute, University of Pennsylvania.

Ray Marshall: professor of economics, The University of Texas, Austin.

Alvin L. Marty: professor, The City College of the City University of New York.

Robert T. Masson: visiting assistant professor, Northwestern University.

Douglas Maxwell: lecturer in economics, State University of New York-Binghamton.

Thomas H. Mayor: associate professor of economics, University of Houston.

Michael B. McElroy: research staff, National Bureau of Economic Research.

David D. McFarland: assistant professor, Graduate School of Business, University of Chicago.

Neville M. Merrett, University of California, Berkeley: assistant professor of economics, California State College, Long Beach.

Laurence H. Meyer: assistant professor of economics, Washington University.

L. Charles Miller, Jr., Fisk University and Vanderbilt University: associate professor of economics, Haverford College.

Leonard S. Miller: assistant professor of economics, State University of New York, Stony Brook.

Harold T. Moody: associate professor of business administration and economics, Clark University.

Donald A. Neilson: assistant professor of economics, Northern Illinois University.

Soren T. Nielsen: research fellow, department of economics and commerce, Simon Fraser University.

Marc Nerlove, Yale University: professor of economics, University of Chicago.

Daniel Newlon: lecturer in economics, State University of New York, Binghamton.

David A. Olson, Smith College: investment analyst, Union Tank Car Corporation, Chicago.

Frederick Dale Orr: assistant professor of economics, University of British Columbia.

James R. Ostas: assistant professor of economics, Bowling Green State University.

Attiat F. Ott, University of Maryland: associate professor of economics, Clark University.

David C. Ott, Council of Economic Advisers: professor of economics, Clark University.

John Owen, Johns Hopkins University: labor economics, New School for Social Research.

Dean O. Popp, Purdue University: assistant professor, San Diego State College.

Richard D. Portes, Oxford University: assistant professor, Princeton University.

B. Michael Pritchett: assistant professor of economics, Brigham Young University.

Usman A. Qureshi: assistant professor of economics, Old Dominion College.

Carl M. Rahm, Columbia University: assistant professor, University of Washington, Seattle.

Helene M. A. Ramanauskas, De Paul University: professor of accounting, Butler University.

John Rapoport: assistant professor of economics, Mount Holyoke College.

George R. Rice, University of Kentucky: assistant professor, Louisiana State University.

Bruce N. Robinson: assistant professor of economics, Haverford College.

Terry L. Roe, Purdue University: assistant professor of agricultural economics, University of Minnesota.

Stephen A. Ross, Harvard University: assistant professor of economics, University of Pennsylvania.

Walt W. Rostow: professor of economics and history, The University of Texas, Austin.

Michael Rothschild, Boston College: assistant professor, Harvard University.

Mahmoud Sakbani: assistant professor of economics, State University of New York, Stony Brook.

James D. Salpietro: assistant professor of economics, State University of New York, Stony Brook.

Warren Sanderson: research staff, National Bureau of Economic Research.

Thomas J. Sargent, Office of Assistant Secretary of Defense: associate professor of economics, University of Pennsylvania.

Donald T. Savage: associate professor of economics, University of Maine.

Gregory K. Schoepfle: assistant professor of economics, State University of New York, Stony Brook.

Nicholas W. Schrock: assistant professor of economics, University of Colorado.

Loren C. Scott, Oklahoma State University: assistant professor, Louisiana State University.

Gerald W. Scully, Ohio University: assistant professor, Southern Illinois University.

Frederick D. Sebold, Boston College: assistant professor, San Diego State College.

Hassam Selim, Colorado University: assistant professor of economics, University of Nebraska, Omaha.

Lawrence Senesh: professor of economics, University of Colorado.

Wayne J. Shafer: assistant professor of economics, Wayne State University.

Donald R. Sherk, Boston College: associate professor, Simmons College.

John Shilling: instructor, Boston College.

Carl S. Shoup: special consultant, National Bureau of Economic Research.

Donald S. Shoup: director of Research Services and Planning, National Bureau of Economic Research.

Ronald A. Smith, Pepperdine College: instructor, department of economics and commerce, Simon Fraser University.

Vinson Snowberger: assistant professor of economics, University of Colorado.

Arthur I. Stoecker: research associate, Iowa State University.

Justin D. Stolen, University of Illinois: assistant professor of economics, University of Nebraska, Omaha.

James A. Storer, Bowdoin College: director, division of economics, department of fisheries, Food and Agricultural Organization, Rome.

Daniel B. Suits: visiting professor of economics, Merrill College, University of California, Santa Cruz.

Paul Sultan, Claremont Graduate School: visiting professor, department of economics and commerce, Simon Fraser University.

Michael Tennenbaum, University of California, Los Angeles: department of economics, California State College, Long Beach.

Richard Tresch: instructor, Boston College.

Keith K. Turner, University of Denver: associate professor of economics, University of Nebraska, Omaha.

Gordon R. Tush, University of Missouri: instructor of economics, University of Nebraska, Omaha.

Michael L. Wachter, Harvard University: assistant professor of economics, University of Pennsylvania.

Charles Waldauer: assistant professor of economics, PMC Colleges.

Robert Wallace: instructor, Boston College.

David C. Warner: assistant professor of economics, Wayne State University.

William Weeks: assistant professor of economics, Wayne State University.

Maurice D. Weinrobe, Cornell University: instructor in economics, Michigan State University.

Richard Weisskoff, Boston College: instructor, Yale University.

Donald A. Wells, Southern Illinois University: associate professor of economics, University of Arizona.

Robert Whitaker, University of Wisconsin: assistant professor, Lafayette College.

C. Glyn Williams, Boston College: associate professor, University of South Carolina.

Donald E. Wise: assistant professor of economics, Clark University.

Donald A. Wittman: acting assistant professor of economics, College Five, University of California, Santa Cruz.

Alan Donald Woodland: assistant professor, of economics, University of British Columbia.

Arthur W. Wright, Oberlin College: assistant professor, University of Massachusetts.

Montague Yudelman: vice-president, OECD Development Centre, Paris.

Frank Zahn: assistant professor of economics, University of Houston.

Dennis Zimmerman: assistant professor of economics, Wayne State University.

Leaves for Special Appointments

William R. Allen, University of California, Los Angeles: visiting professor of economics, Southern Illinois University.

Alan A. Brown, University of Southern California: visiting professor of economics, International Development Research Center, Indiana University.

Arthur F. Burns, senior research staff and honorary chairman of the board of directors, National Bureau of Economic Research: Counselor to the President.

Perry P. Chang, University of Nebraska, Omaha: Asian Development Bank, Manila.

Dean T. Chen, Corporate Planning Services, Deere & Company: visiting associate agricultural economist in Agricultural Experiment Station and Giannini Foundation, University of California, Berkeley.

Francix X. Colaco, Brown University: Economic Policy Fellow, The Brookings Institution.

Herman E. Daly, Louisiana State University: visiting research scholar, Yale University Growth Center.

Gerald F. Fox, Smith College: visiting assistant professor of demography, University of California, Berkeley.

A. Myrick Freeman III, Bowdoin College: visiting scholar, Resources for the Future, Inc.

J. Richard Huber, University of Washington, Seattle: University of the Philippines, Manila.

David Major, The City College of the City University of New York: U. S. Army Corps of Engineers Water Conservation.

Geoffrey H. Moore, senior vice president research, National Bureau of Economic Research: U. S. Commissioner of Labor Statistics.

G. Warren Nutter, University of Virginia: Assistant Secretary of Defense for International Security Affairs.

James Quirk, University of Kansas: visiting professor, University of Washington, Seattle.

Potluri M. Rao, University of Chicago: visiting assistant professor of economics, University of Washington, Seattle.

Gaston V. Rimlinger, Rice University: economic adviser, Ford Foundation, Nigeria.

Sam Rosen, University of New Hampshire: visiting professor of economics, Institute of Social Studies, The Hague.

Paul N. Rosenstein-Rodan, Massachusetts Institute of Technology: visiting professor, The University of Texas, Austin.

Kenneth J. Rothwell, University of New Hampshire: U.N. Development Program, Indonesia.

Harl E. Ryder, Jr., Brown University: Ford Faculty Research Fellow.

Ronald Soligo, Rice University: Ford Faculty Fellowship.

Franklin D. Van Buer, Northern Illinois University: Ministry of Economic Affairs, Ghana.

L. S. Venkataramanan, University of Minnesota: head, Division of Agricultural Economics, Indian Agricultural Research Institute, Pusa, India.

Paul A. Weinstein, University of Maryland: Governor's Executive on Manpower & Labor, Annapolis, Maryland.

Robert E. Willard, University of Texas, El Paso: economist, Federal Reserve Bank of Atlanta.

Resignations

Marcelle Arak, Bernard M. Baruch College of the City University of New York.

Robert W. Doede, University of Pennsylvania, June 1969.

William Hamburger, New School for Social Research.

C. Duncan MacRae, Massachusetts Institute of Technology.

Helen Malenbaum, University of Pennsylvania, June 1969.

George N. Monsma, Jr., Amherst College, June 1969: Calvin College.

Phillip Nelson, New School for Social Research.

John C. Norby, chairman of the department of economics, California State College, Los Angeles.

Richard J. Olsen, University of Massachusetts, January 1969.

Jack Rich, Bernard M. Baruch College of the City University of New York.

Peter Schulkin, Boston College, June 1969: economist, Federal Reserve Bank of Boston.

Bernard Seligman, Bernard M. Baruch College of the City University of New York.

Franklin B. Sherwood, University of Massachusetts, January 1969.

William J. Stober, Louisiana State University: September 1969, University of Kentucky.

Sang C. Suh, Clark University: International Bank for Reconstruction and Development.

Allen V. Wiley, Bowling Green State University.

The American Economic Review

The American Economic Review

Vol. LIX, No. 4 Part 2 September, 1969

SUPPLEMENT

SURVEYS OF NATIONAL ECONOMIC POLICY ISSUES AND POLICY RESEARCH

Price $1.00

Foreword

The two surveys of national economic policy issues and policy research published in this Supplement to *The American Economic Review* are the second in a series of studies of economics in foreign countries, commissioned by the Publications Committee of the American Economic Association and under the editorship of Harry G. Johnson. The previous series, edited by George W. Hildebrand, sought to acquaint Anglophone economists with the significant developments in economics since the Second World War in the national literatures of the major non-English-speaking countries. This series, by contrast, is directed at economic policy issues, and the controversies and research to which they have given rise, in countries selected for one or more of three reasons: the intrinsic interest of the policy issues, the relevance of the policy issues to current policy issues in the United States, and the interest of the countries themselves as areas of involvement of American foreign economic policy with which U. S. economists are likely to become concerned.

HARRY G. JOHNSON

Contributions to Indian Economic Analysis: A Survey

By JAGDISH N. BHAGWATI AND SUKHAMOY CHAKRAVARTY*

TABLE OF CONTENTS

Contributions to Indian Economic Analysis: A Survey

By JAGDISH N. BHAGWATI AND SUKHAMOY CHAKRAVARTY*

Any survey of contributions to economic analysis in India, even though confined to the post-war years and to issues arising from domestic economic events and policy, runs into exceptional difficulties. Not only has practically every conceivable problem been raised and discussed by economists, in a country where interest in economic issues dates back at least to the latter half of the 19th century;[1] but there have also been numerous committees and commissions whose report have led to a voluminous literature.

Ruthless selectivity has thus been inevitable. We have generally focussed, in this survey, on contributions which meet the following criteria: (1) they should have analytical interest, either theoretical or empirical; (2) they should be made by Indian or India-based economists; and (3) they should have some bearing on Indian economic policy issues, even though they cannot necessarily be demonstrated to have arisen in consequence thereof or to have had any impact on policymaking.

The Survey thus rules out of consideration the vast bulk of official literature, whose analytical base is frequently largely minimal, as also the purely descriptive and institutional material from non-official sources (such as the Indian Statistical Institute) whose utility otherwise is not to be minimized. Equally, the Survey does not extend to the growing numbers of contributions to general theoretical economic analysis that Indian economists have begun to make, as is evident from the contents of reputed journals in the last decade.[2]

This Survey, therefore, is neither a comprehensive account of the state of econ-

* The authors are respectively Professor of Economics, M.I.T. and Professor of Mathematical Economics, Delhi University

This Survey was commissioned by the American Economic Association. We have profited from discussions with, or comments on early drafts from, numerous colleagues and friends, including V. K. Ramaswami, Alan Manne, K. N. Raj, Dharm Narain, A. N. Khusro, C. H. Hanumantha Rao, V. M. Dandekar, Padma Desai, P. N. Dhar, Ashok Desai, B. S. Minhas, V. R. Panchmukhi, T. E. Weisskopf, Richard Eckaus, Paul Rosenstein-Rodan, Harry G. Johnson, T. W. Schultz and T. N. Srinivasan. We wish to thank K. Sundaram for assistance in compiling the bibliography and for making valuable suggestions. While the bulk of the *detailed* work in Part I was undertaken by Chakravarty and in Parts II and III by Bhagwati, the Survey has been planned and also written jointly by both authors. We may also add that the Survey was finished essentially by May 1968 and, in consequence, we have not been able to cover some additional material of interest which has become available since.

[1] For example, the prominent participants in the Independence movement paid considerable attention to economic issues; and India produced her own brand of List-type arguments in favor of departure from free trade in India's national interest [118] [128] [141]. The Indian Economic Association was founded in 1918 and the Indian Society of Agricultural Economics in 1939. Among the important centers of economic research today are the three Centres of Advanced Study in Economics, recognized and financially supported as such by the University Grants Commission: Bombay University, Gokhale Institute of Politics and Economics (Poona), and Delhi School of Economics; and at least two other institutions: the Institute of Economic Growth (Delhi) and the Indian Statistical Institute (New Delhi).

[2] These contributions primarily range over the theory of growth and planning (e.g. S. Chakravarty, G. Mathur, A. K. Sen and T. N. Srinivasan), international trade theory (e.g. P. K. Bardhan, J. Bhagwati, B. S. Minhas, V. K. Ramaswami and T. N. Srinivasan), and econometric theory (e.g. A. L. Nagar). The trend is firmly established and strongly upward.

omic research in the country[3] nor does it pretend to give an exhaustive picture of the policy issues that have been discussed on the Indian scene since 1947 when India gained independence.

The Survey is broadly divided into three areas: (1) planning theory and techniques; (2) agriculture; and (3) foreign trade. The vast majority of India's policy issues, and analytical literature, fall within one or more of these categories. With her programs for economic development, initiated through the First Five–Year Plan (1951/1956) and continued through two successive Five–Year Plans, the question of overall Plan formulation, and investment criteria in particular, has engaged the attention of many economists. Largely because agriculture is the overwhelmingly important economic activity in the economy, and its capacity to act as a significant brake on growth via its role as the supplier of wage goods to other sectors has been increasingly appreciated, this sector has also attracted considerable economic analysis. And finally, the foreign trade sector has been the focus of interesting debate. The questions raised by foreign aid, foreign investment, the import control regime and export subsidization have led to insights of wider interest.

Since the overall Plan formulation literature inevitably embraces some of the questions raised by agricultural and foreign trade policies, our discussion begins with the survey of the planning literature and only then proceeds to an evaluation of the literature that concerns itself with the *remaining* issues in the areas of agricultural and trade policies.

I. *Planning Theory and Techniques*

The formulation of the successive Five–Year Plans in India has led to a steady evolution of economic thinking on questions relating to planning theory and techniques. As we shall soon argue, however, the interplay between Plans and economic thinking has often been tenuous. At times there may even have been post-facto rationalization of investment decisions taken on political grounds by ingenious designing of suitable models. At other times, model–building and analysis have inevitably gone ahead of the Plans. However, it is possible to identify with each Plan certain basic model–types which have provided the intellectual backbone to that Plan and were the object of extensive economic debate.

Thus the First Five–Year Plan, which was essentially a collection of several projects, contained at the same time a Harrod-Domar type exercise which sought to examine the growth rates that would be achieved by specification of the (feasible) marginal savings rate and the resulting average savings ratio. The Second Five–Year Plan, on the other hand, marked a distinct departure in favor of the Feldman–Mahalanobis type of structural model which emphasises the *physical* aspect of investment and thus leads, subject to certain restrictive assumptions about transformation possibilities domestically and through foreign trade, to the proposition that raising the rate of investment requires increased domestic manufacture of capital goods. This shift from a Keynesian, "flow" analysis which emphasised the necessity to raise savings (and hence implicitly assumed that the savings could be transformed into required investment) to a "structuralist" view which emphasised the transformation constraint and the supply of capital goods to sustain growing investment (while implicitly assuming that the system would generate the savings to "finance" the growing supply of investment goods) was the most dramatic episode in the evolution of planning literature and debate in India. The formulation of

[3] Among useful attempts in this direction, with respect to agriculture, see Gupta [61] and Dandekar [28].

the Third Five-Year Plan, by contrast, marked a shift away from these *simple decision-models*: the achievement of inter-industrial consistency was attempted in some detail this time. As we shall soon see, the shift to interindustrial exercises not only underlay plan formulation but was also the characteristic of planning exercises undertaken by economists and teams associated with the Indian Planning Commission. These *multi-sectoral models* were also characterised by their explicit extension to questions of intertemporal choice: questions which had been raised as early as 1955 by Ragnar Frisch when he visited the Indian Statistical Institute which was the intellectual center for formulation of the Second Plan (1956/1961).

Having identified synoptically the main outlines of shift in planning techniques in India through the three Plans, we now proceed to survey the major ideas and contributions in this area, considering each of the three planning periods in turn.

The First Five-Year Plan

The first identifiable planning model used in India was developed by the authors of the First Five-Year Plan document which the Government of India placed before the country in 1951.[4] The model was not given an explicit analytical form, but was implicit in the numerical figures which constituted the perspective plan for developing the Indian economy [179, Chapter I]. It was essentially a simple variant of the growth model associated with the names of Harrod and Domar. The sole modification, but a crucial one, was the distinction between the average and the marginal propensities to save. The capital-output ratio was assumed to be the same on the margin as on the average. No gestation lags were introduced. The model was developed for a closed economy (although it can naturally be easily extended to deal with an open economy, with one part of investment being financed by import surplus). The basic equations underlying the growth process were the following:

$$(1) \qquad I_t = S_t;$$

$$(2) \qquad S_t = aY_t - b;$$

$$(3) \qquad Y_t = \alpha K_t$$

$$(4) \qquad I_t = \dot{K}_t$$

Here I_t stands for investment at 't', S_t for the corresponding amount of savings, Y_t stands for income. All the equations excepting (2) are the same as in the Harrod-Domar model. Equation (2) introduces the distinction between marginal and average propensities to save. The model leads tothe basic differential equation $\dot{K}_t = a\alpha K_t - b$ which can be easily solved to give us the time profile of capital stock and output. We get:

$$(5) \qquad K_t = (K_0 - b/a\alpha)e^{a\alpha t} + b/a\alpha.$$

Notice that unlike the usual Harrod-Domar model, the rate of growth here rises from period to period (provided of course $a > S_0/Y_0$). Thus an economy which decides to save more on the margin than on the average can hope to do better and better over time in terms of its rate of growth. The asymptotic relative rate of growth of the system is given by the expression $a\alpha$.[5]

Such Harrod-Domar type models have been explicitly used, with considerable advantage, as the framework for plan formulation in other countries (e.g. by Jan Tinbergen for the first Turkish Plan). They are useful in indicating the basic

[4] There *had* been earlier attempts at putting together "plans" for India: e.g., the Bombay Plan [169] in 1944. However, no conceptual framework, in terms of an explicit or implicit planning model, underlay any of these exercises.

[5] For any specific t, $r_t < a\alpha$ where r_t is the relative rate of growth of income at time t.

macro-economic features that any more elaborate construct would equally have to satisfy. Further, they have served as "simple" mechanisms for computing the external assistance that may be necessary for supplementing domestic savings to sustain projected growth rates in income.[6]

However, such a Harrod–Domar model obscures problems of importance. For example, concentration on the flow equilibrium, and the implicit assumption that there are no "structural" difficulties in transforming savings into (desired forms of) investment may ignore real constraints in the economy. Further, even within the framework of its assumptions, the model ignores the fundamental choice problem of planning over time, which requires a weighing of present versus future gains, by assuming a constant marginal propensity to save for the economy.

The connection between the actual First Five–Year Plan and the Harrod–Domar type model contained in the document was left vague by the planners. It appears as though the selection of projects for governmental expenditure reflected essentially the "overhead–capital" approach to developmental planning and the model was largely an intellectual appendage with little impact on actual Plan formulation, although it did serve to give some kind of longrun perspective to the Plan.

The Second Five–Year Plan

By contrast, the Second Plan pattern of industrial investment, with its marked shift in favour of capital goods industries, was deeply influenced by the two-sector growth model developed by P. C. Mahalanobis [95].[7] This model was independently developed by Feldman in the Soviet Union in the 1920s and later revived by Domar [47] in a considerably improved form. The basic model, as stated by Mahalanobis, can be described briefly.

Current investment flow I_t is divided into two parts, $\lambda_k I_t$ and $\lambda_c I_t$, where λ_k indicates the proportion going to the capital goods sector and λ_c the corresponding proportion for the consumption sector.

It is clear that

$$I_t - I_{t-1} = \lambda_k \beta_k I_{t-1} \tag{7}$$

and

$$C_t - C_{t-1} = \lambda_c \beta_c I_{t-1}. \tag{8}$$

Now the first equation implies that

$$I_t = I_0(1 + \lambda_k \beta_k)^t. \tag{9}$$

Further, $C_t - C_0$ can be written as

$$\sum_{\tau=1}^{t} (C_\tau - C_{\tau-1}) = \sum_{\tau=1}^{t} \lambda_c \beta_c I_{\tau-1} \tag{10}$$

$$= \lambda_c \beta_c I_0 + \lambda_c \beta_c I_1 + \cdots + \lambda_c \beta_c I_{t-1} \tag{11}$$

$$\begin{aligned} &= \lambda_c \beta_c I_0 + \lambda_c \beta_c I_0 (1 + \lambda_k \beta_k). \\ &\quad + \lambda_c \beta_c I_0 (1 + \lambda_k \beta_k)^{t-1} \end{aligned} \tag{12}$$

$$= \frac{\beta_c \lambda_c}{\beta_k \lambda_k} I_0 [(1 + \lambda_k \beta_k)^t - 1] \tag{13}$$

Since $I_t - I_0 = I_0\{(1+\lambda_k\beta_k)^t - 1\}$, we get by adding it to $C_t - C_0$ in the preceding equation:

[6] Suppose the planners are ambitious enough to set a target rate of growth in income which implies an investment rate in excess of the current savings rate. In an open economy, this would not raise any problem so long as the required amount of foreign aid (F_t) is available to meet the domestic resource gap. However, if the growth process is of the type described in the text (where the economy saves more on the margin than on the average) then the required amount of foreign aid would diminish from year to year, provided the growth rate is kept constant. The time t^* for which F_t vanishes may be defined as the time of attainment of self-sustained growth. This value for t^* may be compared with the value t^{**} for which the economy would reach the desired growth rate left to itself. The difference between t^{**} and t^* may be used to give one measure of the beneficial influence of foreign aid on economic growth. Thus, the simple growth process described in the text can help one to obtain answers to questions relating to the volume of external assistance that is necessary.

[7] Numerous specific criticisms of the analysis of the Mahalanobis model were made at the time, among them being Chakravarty [20], Tsuru [172] and Mitra [117].

$$(14)\quad C_t - C_0 = \frac{\beta_c\lambda_c}{\beta_k\lambda_k} I_0\{(1+\lambda_k\beta_k)^t - 1\},$$

the complete solution for output at time t, where

$$(15)\quad Y_t = Y_0\left[1 + \alpha_0\left(\frac{\beta_c\lambda_c + \beta_k\lambda_k}{\beta_k\lambda_k}\right) \cdot \{(1+\lambda_k\beta_k)^t - 1\}\right]$$

where $\alpha_0 = I_0/Y_0$, the initial investment–income ratio.

Several things are quite clear from this equation. First we note that the relative rate of growth of consumption or output is changing over time. It is also clear that the asymptotic rate of growth of the system is given by $\lambda_k\beta_k$ where λ_k is the crucial allocation ratio which indicates the proportion of capital goods output which is devoted to the further production of capital goods. Thus a higher λ_k would always have a favourable effect on the asymptotic growth rate of the system, irrespective of whether it is consumption or output. But what about its immediate effect on consumption? If $\beta_c > \beta_k$, then a higher value of λ_k would imply a lower immediate increment in consumption. Thus, there is implicit in the choice of 'λ_k' a choice of alternative time streams of consumption.[8]

It may be further noted that, while the implicit assumption underlying the aggregative model discussed earlier was that the savings rate was a reflection of the behavioral characteristics of the decision-making units such as the household, the corporate sector or the government, Mahalanobis effectively made it a rigid function of certain "structural" features such as the capacity of the domestic capital goods industry and capital–output ratios of the capital goods sector and the consumer goods sector. By making the allocation ratio of current investment going into investment goods sector the policy variable, he showed that a higher allocation would mean a higher saving rate on the margin and hence a greater rate of growth of output or consumption.[9] This can be seen readily by noting that $\lambda_k\beta_k/(\lambda_c\beta_c + \lambda_k\beta_k)$ is none other than the share of incremental investment in incremental output. Macroeconomic balancing for a closed economy would then imply that this is also the share of incremental savings in incremental income. If $\beta_k = \beta_c$, then this ratio of incremental savings to incremental income is exactly equal to λ_k. If $\beta_k \neq \beta_c$, then $\Delta I/\Delta Y$ is a more general function of λ_k, β_k, β_c but the fundamental qualitative point remains unaltered.

Despite the fact that the Mahalanobis model is a severely rigid construct, it has one important virtue. This lies in its recognition of the fact that capital equipment once installed in any specific producing sector of the economy may not be shiftable.[10] An important consequence is that changes in the savings rate, and hence in the rate of investment, are not necessarily feasible and become conditional upon the composition of the existing capital stock; hence, optimal programs of

[8] Mahalanobis did not address himself to the question of how to resolve this choice problem. He, however, pointed out that a specification of the horizon over which the planning was done was essential if any meaningful answer is to be given to the choice of λ_k. This is undoubtedly correct but, as more recent analysis dealing with this question has shown, specification of a planning horizon is only one of the many prerequisites for choosing an optimal path of development over time. We have to make some assumption regarding the nature of the intertemporal utility function as well as the terminal conditions of the problem. We shall deal with these questions to a certain extent when we come to the discussion of the more recent planning models constructed in the Indian context.

[9] In deriving this cental proposition, Mahalanobis implicitly ignored the role of foreign trade altogether and assumed that the government was in a position to control consumption completely.

[10] Whether, however, the Mahalanobis-assumed non-shiftability from consumer goods to investment goods capital equipment is greater than that within the former group, and how important it is anyway, are matters on which *evidence* is scant and, as we shall soon argue, was in any case not sought by the Indian planners before adopting Mahalanobis' ideas.

capital accumulation worked out under the assumption of nonshiftability differ crucially from those derived from models with complete shiftability in capital stock among alternative uses.[11]

It needs to be stressed, of course, that foreign trade also can get the economy out of the problems raised by limited transformation possibilities domestically owing to nonshiftability of capital equipment: the assumption of a closed economy automatically rules out this important escape route from the problems raised by nonshiftability. Of course to escape these problems completely, we would have to assume the possibility of indefinite transformation at constant rates—the so-called "small country" assumption in trade theory. Therefore, the essential problems raised by nonshiftability will persist if the reciprocal foreign demands facing the planning country are less than perfectly elastic.

Mahalanobis, who assumed a closed economy and total nonshiftability of the capital stock from the consumption goods to the investment goods sector, appears to have used his model merely to provide the rationale for a shift in industrial investments towards building up a capital goods base. However, the *precise* choice of the proportion of investments in the capital goods sector, during the Second Plan and possibly, thereafter, appears to have been arbitrary—at any rate, if there were specific economic considerations underlying it, these were not spelled out. In any case, an optimal choice thereof would have required, at the minimum, a quantification of the transformation constraints (both domestic and foreign)—and we know that neither was attempted.

Indeed, it appears quite plausible to argue that Mahalanobis (who had just then visited the socialist countries and with whose economists he had close contacts) was impressed with Soviet thinking on industrialization, with its emphasis on the building-up of the capital goods base, without full recognition of the fact that such a strategy presupposes constraints on domestic and foreign transformation which need to be empirically verified. Further, it seems likely that, being a physicist by training and a statistician by practice, he directly identified increased investment with increased availability of capital goods, which in turn he identified with domestic production thereof, ignoring foreign trade in particular.[12] It is interesting that the Second Plan did not explicitly state the rationale of the shift to heavy industries in terms of foreign trade constraints, so that the later justification of this strategy by alluding to "stagnant world demand" for Indian exports comes somewhat close to a *ex post facto* rationalization. Indeed, the Second Plan's examination of export earnings through the Plan is so cursory that it is difficult to believe that the "stagnant world demand for Indian exports" assumption, by virtue of which the shift to heavy industries was later sought to be justified, was seriously made: such a *crucial* assumption, if made, would surely have been examined more intensively! Further it is important to note that the preceding Five–Year Plan's experience with the balance of payments *and* exports was comfortable, so that it hardly seems likely that the export prospects could have been viewed with such pessimism as has later been imagined.[13]

While, therefore, the Mahalanobis two-sector model was used to provide the rationale for a general shift in investments to building up a capital goods base,

[11] Cf. Chakravarty [22] for a detailed treatment of the analytical issues raised by planning for optimality in the context of models with non–shiftable capital.

[12] This probably accounts for his model [96] in the *Draft Frame* of the Second Plan taking no *explicit* account of savings, whereas *economists* looking at growth inevitably started from the savings end.

[13] These arguments have been developed more fully in Bhagwati and Padma Desai [14].

though the actual magnitude of this shift was otherwise determined, Mahalanobis provided yet another model, a four–sector model [96], which broke down total investment among three further sectors, in addition to the capital goods sector: (1) factory production of consumer goods; (2) household production of consumer goods, including agriculture, and (3) the sector providing services such as health, education etc.[14]

Mahalanobis assumed that all four sectors had independent output–capital and labor–output ratios. These were symbolized by β_1, β_2, β_3, β_4, θ_1, θ_2, θ_3 and θ_4 respectively. He assumed a given total of investment. The problem was to allocate the total between the sectors in such a way that specified increases in income (ΔY) and in employment (ΔN) were reached. The policy variables were the shares of investment going to each sector, denoted by λ_1, λ_2, λ_3 and λ_4.

The model was determined, of course, only if one of the three independent λ's (the policy instruments) was exogenously determined, since there were only two objectives: ΔY and ΔN. With the λ for the capital goods sector given a pre-assigned value (the reason for which was never spelled out clearly) the system was solved by Mahalanobis to assign investments among the three remaining sectors. However, as Komiya [78] pointedly noted, the Mahalanobis solution was inefficient, in that it was situated in the interior of the feasibility locus between incremental output and incremental employment. Thus, greater employment and/or output could have been obtained by merely reallocating the given investments among the three sectors, although such a solution would not assign a positive fraction of investment to every sector.

[14] The entire economy was supposed to be divided into these four sectors.

The very fact that a simple linear programming exercise by an outsider could show the inefficiency of the Mahalanobis allocations, in conjunction with the fact that Mahalanobis did not use this technique even though the planners at his Indian Statistical Institute were certainly not lacking in knowledge of these elementary techniques,[15] indicates that the four–sector model was essentially produced to impart (unsuccessfully, as it turned out) intellectual respectability to investment allocations arrived at on other, unspecified considerations. This conclusion seems also warranted by the fact that the statistical source of the parameters (relating to labor–output and capital–output ratios) was not spelled out. Nor was any attempt made to reconcile the model with the real facts of the situation, especially the presence of foreign trade.

The very limitations of the Mahalanobis two–sector and four–sector models pointed to the need for a more extensive, multi–sectoral and multi–period model for more efficient resolution of the choice problems facing the Indian economy. Such models were to be constructed during the Third Plan period and we shall go on to discuss them. However, it is pertinent to mention here an alternative approach to Indian planning, rival to that of Mahalanobis, which was put forward at the time of the Second Plan formulation by P. R. Brahmanand and C. N. Vakil [19].

Their approach constitutes in some ways the polar opposite of the position taken by Mahalanobis. While the latter's whole emphasis was on the role of fixed capital, Vakil and Brahmanand's entire emphasis was on the role of wage goods as

[15] The Indian Statistical Institute is internationally renowned for its contributions to mathematical statistics and its distinguished Faculty which currently includes two Fellows of the Royal Society in this subject. Besides, they had the benefit of visits, at the time, by Richard Goodwin, Jan Tinbergen, Ragnar Frisch and Oskar Lange.

capital. This approach was therefore related to the Marxian concept of variable capital, since in common with Marx they assumed that the wages were paid as advances in the beginning of the production period. However, the operational part of this approach was derived from the assumption that there existed massive overpopulation in agriculture. They did not subject the concept of disguised unemployment in agriculture to any critical investigation, nor did they try to measure the extent of such disguised unemployment in the Indian context. They were following the tradition set by P. N. Rosenstein–Rodan and R. Nurkse in taking it as obvious that a massive reserve army of labor existed in rural areas.[16] Further they assumed, as Nurkse did, that the disguised unemployed must possess considerable savings potential since labor could be transferred from agriculture without lowering production and kept at work producing real capital goods by the payment of wages consisting exclusively of food.

Several assumptions were made by Vakil and Brahmanand to formalize their system. First, the wage good was assumed to be exclusively food. Second, it was assumed that labor could produce capital goods without the assistance of other factors of production. Thirdly, they assumed that a mechanism existed by which average consumption on the farm could be kept constant subsequent to the transfer of labor, and the transferred labor's consumption on the farm be siphoned off to feed it while it was engaged in producing capital goods. To be sure, they recognized the possibility of leakage in this connection. They assumed that the whole of the hypothetical surplus might not be procurable and further that there might be a need to provide for slightly higher consumption levels than in agriculture to workers engaged in construction. Thus they derived a "multiplier" formula in which the multiplicand was an autonomous increase in the stock of wage goods in the hands of the planning agency.[17]

However, they did not pay any attention to the possibility that the production process may not be of the simple Austrian type that they had assumed: labor→capital goods→consumer goods. If the facts of life dictated a circular model of the sort analyzed by Marx (in the second volume of *Das Kapital*) and more recently by Leontief and others, the mere availability of labor alone would not solve the problem of greater capital formation. Extra capital equipment would be needed and would need to be either produced at home or imported, hence raising the complex of issues raised earlier in connection with the Mahalanobis model. As regards the composition of wage goods, if nonfood items were necessary and if there were no corresponding excess capacity in the domestic consumer goods industries, then once again the creation of extra capacity and its synchronization with the deployment of extra labor would be involved, thus necessitating a more elaborate approach.

[16] The early writings of Brahmanand [18] in fact anticipated the notion of disguised unemployment, although the concept was not fully developed. Brahmanand's interest in this question had arisen from a general examination he was undertaking of the applicability of the Keynesian theory to the economic situation in India. In this connection, V. K. R. V. Rao's [143] analysis of the Keynesian multiplier in the Indian situation is also of interest.

[17] The Vakil–Brahmanand approach could be readily used to provide a rationale for the policy of importing food as a means to step up capital formation and generate extra employment. It is interesting that these authors did not make any effort to link up their analysis directly with the question of food aid. Although such an approach was implicit in the early work of Dandekar [27] and others, it was only much later that Chakravarty and Rosenstein–Rodan [26] tried to develop the logic of food aid somewhat more fully in an analysis which was considerably influenced by a model similar to that provided by an economy with massive rural overpopulation.

The social welfare function implicit in Vakil and Brahmanand's approach to planning was novel at the time: they were emphasizing the need to minimize the time needed to reach full employment. Whether such an objective constitutes an adequate social welfare function is very doubtful;[18] but that it is an objective which needs very careful consideration is beyond doubt. Mahalanobis' model, at least in its formal aspects, had paid no attention to this.

Despite its limitations, Vakil and Brahmanand's attempts to build an analytical scheme which tried to tackle the problem posed by disguised unemployment in agriculture deserve emphasis.[19] In discussions subsequent to the formulation of the Second Plan, the wage goods approach has not figured prominently. Amongst eminent economists, Gadgil [55] has been the only one to draw pointed attention to the importance of mobilizing rural labor to build social overhead capital, which he did as late as 1961. In more attenuated forms, however, this aspect of the planning problem is still alive.

The Third Five-Year Plan

The Third Plan was not entirely pioneering in its attempted utilization of multi-sector balances to achieve consistency. Quite aside from Ragnar Frisch's suggestions in this respect [54], Jan Sandee [150] had actually constructed a simple linear programming model during his visit to the Indian Statistical Institute during 1957/1958, which was used to maximize aggregate consumption in a terminal year (1970) as an excess of consumption over a base year (1960), subject to maintaining intersectoral consistency conditions and feasibility conditions on the side of the balance of payments.

While the Sandee model was essentially a straightforward, *static* linear programming exercise, it had one analytically important feature which deserves special mention. This relates to his treatment of investment in the terminal year. If this year were taken really to be the terminal year of the Plan, then clearly there is no justification for having any investment at all in that year (assuming, of course, that all investments fructify beyond the one year horizon). Since no planner ever takes such a myopic point of view, it is necessary to elaborate a rationale for introducing investment activity in single-period optimization models. One can theoretically conceive of several procedures which can be used for this purpose. Sandee's procedure was to assume that, over the decade of 1960–70, investment flows should increase linearly every year. We may spell this out.

Let us denote the year 1960 by '0' and the year 1970 by T. Then the cumulated investment over the period is given by

$$\int_0^T I(t)dt.$$

Sandee assumed $I(t)=a+bt$. Then we have

$$\int_0^T (a+bt)dt = aT + \tfrac{1}{2}bT^2.$$

The proportion of total investment that must take place in the T-th year out of the total over the entire T year period is given by $a+bT/aT+\frac{1}{2}bT^2$. Now applying this factor to the investment demand for the product of the i-th sector induced by output increase in 1970 over 1960, which equals

$$\sum_j b_{ij}\Delta X_j,$$

we get the estimate for investment of the i-th type for the year for every i. Hence the vector of goods to be delivered on the investment account in the terminal year is determined. Intersectoral deliveries on the

[18] See Chakravarty [23] on this issue.

[19] On the question *whether* disguised unemployment exists, we survey the Indian literature in Section II.

current account, together with the balance of payments considerations, must be included. Sandee, then, proceeded to maximize consumption in 1970 subject to obeying a lower limit on total investment summed over all the sectors and a few other inequality constraints.[20]

The work underlying the Third Plan was nowhere as explicitly set out as in Sandee's exercise, although balances of supply and demand at a detailed sectoral level were set out. Reddaway [145], who was associated with the Perspective Planning Division of the Indian Planning Commission,[21] undertook a systematic supply-demand balance exercise, for many industries, essentially putting together different target outputs, imports and demands to test for simple consistency for the year 1965–66, the terminal year of the Third Plan. If this exercise were construed as constituting simple, *ad hoc* checks on targets supplied to Reddaway by the Perspective Planning Division (PPD) it was valuable. But construed as an attempt at devising a full-fledged Third Plan, with only partial targets of production supplied by the PPD, the exercise was less satisfactory even within the framework of testing for consistency (as distinct from optimality). This was pointed out by Padma Desai [42] who, on attempting to formalize the Reddaway exercise, found it underdetermined despite efforts at discovering (from Reddaway's work) ways in which the model might have been intended to be closed.[22] The problem clearly arose from Reddaway's omission to state his model, if any, in formal terms. It therefore throws into focus the need for stating carefully the model underlying the investment allocations and related decisions, quite aside from the theoretical elegance and scrutiny of otherwise vague assumptions that such a procedure would entail.

In fact, during the Third Plan period itself, many economists such as Alan Manne, Ashok Rudra, Sukhamoy Chakravarty, Richard Eckaus, Louis Lefeber and Kirit Parikh, turned to precisely this kind of work in connection primarily with the Fourth Plan: whereas Manne and Rudra were to build static, multi-sector consistency models, the Chakravarty–Eckaus–Lefeber–Parikh exercises were to be concerned with explicitly dynamic, multi-sector models.[23]

Before we discuss either of these two developments, both of which marked improvements over the earlier computational models of planning, we should note that the Third Plan not only marked a shift to examination of consistency at the inter-sectoral level but also incorporated some fresh, though embryonic, analytical thinking. The notion that foreign trade might be the bottleneck to increasing the rate of investment had come more sharply into focus, instead of being the implicit premise of a Mahalanobis type of investment strategy.[24] Thus, significantly more than

[20] It may be noted that, as Sandee assumed that the balance of payments in 1970 would be such as to require no import surplus, he could pay no attention to any parametric variation in foreign aid availability, a question which has been repeatedly posed by many model builders in India since then.

[21] Reddaway was on the M.I.T. Center for International Studies Program, under which several distinguished economists were associated with the Planning Commission's work. P. N. Rosenstein-Rodan and Max Millikan headed this program, which brought to India numerous economists including I. M. D. Little, Trevor Swan, Arnold Harberger, Louis Lefeber, J. Mirrlees, Richard Eckaus, and Sir Donald MacDougall.

[22] See the interchange between Padma Desai and Reddaway [146] on this issue.

[23] As it eventually turned out, the Fourth Plan was deferred by three years, largely thanks to the dislocation of aid flows following the Indo-Pakistani war of late 1965 and two unprecedented agricultural droughts in 1965–66 and 1966–67 which, in turn, led to a recession in industrial investments. The planning exercises, which had not anticipated these major disturbances, turned out to be irrelevant to the immediate situation. Whether, however, the Fourth Plan should have been postponed in consequence is a matter on which there has been much debate.

[24] This view was, at least partly, a reflection of the stagnation in India's export earnings during the decade

with the Second Plan, the investment decisions in the Third Plan were taken with an explicit attention to the role of foreign aid in breaking this bottleneck and the possible desirability of "using aid to end aid," and reach self-sustained growth at some foreseeable future date. To put it differently, the Third Plan investments, which continued the shift to the heavy industrial sector, were taken against the notion that foreign aid would enable the economy, by permitting these investments, to cross over the hump (posed by the foreign trade constraints) from a low growth rate equilibrium to a high growth rate equilibrium. This precise view was to be the basis of more formal exercises by Manne and Bergsman [100] in connection with the Fourth Plan work.

Models for the Fourth Plan

Among the detailed, *static* exercises attempted during the work on the Fourth Plan was that by Manne and Rudra [101], who were both working in collaboration with the PPD (which had put out its own "bluebook" of projections based on similar thinking).

Although their exercise was of the standard type, and related to the consistency of the terminal year of the Fourth Plan, there were certain interesting features. For example, they followed Sandee in attempting to give a rationale for investment activity in the terminal year. However, unlike Sandee, they assumed that investment would rise exponentially over the intervening years and hence the proportion of investment to be completed in the terminal year was given by

1951–1960, which made the hypothesis of a foreign exchange bottleneck to Indian development seem much more plausible. Manmohan Singh [163] was later to show that this stagnation was largely a result of domestic policies, although demand factors would have constrained significant expansion of export earnings.

$$\gamma = \frac{e^{rT}}{\frac{1}{r}\left[e^{rT} - 1\right]} = \frac{r}{1 - e^{-rT}} \tag{16}$$

This proportion "γ" was called by them the stock–flow conversion factor. Having done this, they took the consumption vector to be given and tried to find out the gross production vector that would be needed if all the end use activities were to be satisfied at given levels, subject to the assumption of a stipulated stock flow conversion factor. Their procedure may be summarized as follows.

Let $\hat{x}$ stand for the vector of production levels in 1970 and x^0 for the vector of production levels in 1960. Then we get

$$\hat{x}_i + M_i = \sum a_{ij}\hat{x}_j + F_i + \gamma \sum b_{ij}(\hat{x}_j - \overset{0}{x}_j) \tag{17}$$

where M stands for imports, F for final demand and a_{ij} and b_{ij} are the standard current input–output and capital coefficients respectively. If M were written as $[m]\hat{x}$ where $[m]$ is a diagonal matrix of import coefficients, then we may write the solution of the above as:[25]

$$\hat{x} = (I + m - A - \text{`}\gamma\text{'}B)^{-1}[F - \gamma B x^0] \tag{18}$$

where I is the unit matrix.

Since Manne and Rudra were concerned with constructing a terminal year model for the Indian economy, they did not specify in a complete way the path the economy was to follow from a given initial

[25] It is clear that the choice of 'γ' cannot be completely unrestricted. It must satisfy an upper bound restriction if $(I+m-A-\gamma B)^{-1}$ is to be non negative. From the theorems on non negative square matrices we know that this would be the case provided m were sufficiently small so that $(A+\gamma B-m)$ is non negative, and $1>r_0$ where r_0 is the Frobenius root of the matrix $(A+\gamma B-M)$. A sufficient condition for this to hold is that $\max_j \Sigma_i(m_{ij}+a_{ij}+rb_{ij})<1$ for any suitable choice of units. Hence, given the coefficients of production, the capital coefficients and a given pattern of competitive import demand, γ must have at least an upper bound.

point to the terminal configuration. Thus this model may be said to give a perspective rather than a plan. However, the construction of a terminal configuration is an essential ingredient in any finite horizon planning model so that the Manne–Rudra model did provide some guidelines for the planner, even though it did not specify a complete time phased course of action.

Models which give some guidelines regarding the phasing of investment can be usefully divided into two categories.[26] One of these categories may be described as giving us intertemporally consistent planning models. The other category, somewhat more ambitious in scope, deals with optimization over time, and has been experimented with by Chakravarty, Eckaus, Lefeber and Parikh.

Chakravarty and Eckaus [24] first outlined the logic of an intertemporally consistent multisectoral planning model and pointed out the basic difficulties. They did not, however, compute numerical growth paths. It is of some analytical interest to give a brief summary of their arguments.

For the sake of simplicity, let us ignore temporarily all the non-homogeneous elements of a Leontief dynamic system other than consumption. Then, we have a dynamic system characterized by the following equation:

$$X(t) = AX(t) + B\dot{X}(t) + C(t) \tag{19}$$

where $X(t)$ stands for the vector of output levels, $\dot{X}(t)$ for its rate of change, $C(t)$ for the vector of final consumption, and A and B are the standard Leontief matrices. The complete solution of the dynamic system assuming $C(t)$ to be growing exponentially over time is given,[27] using the notation of a matrix exponential, by the following expression:

$$\begin{aligned} X(t) = {} & {}_e[B^{-1}(I - A)^{+}] \\ & + (I - A - rB)^{-1}Ce^{rt} \end{aligned} \tag{20}$$

where P is a vector of arbitrary constants and C represents exogenous levels of consumption to be determined through policy considerations. The advantage of using this matrix exponential notation is that it shows the analogy with the ordinary scalar case involving one variable quite explicitly.

The above equation holds provided B has an inverse, a condition which is often not satisfied. The solution for more general time paths of $C(t)$ may also be worked out, either in an analytical form or through numerical approximations.

Now if $X(0)$ and $X(T)$ are both given, in the first case from history and in the second case from the planner's specifications, then we have $2n$ equations to determine the unknown P and C's. This is the logic of the consistency models. However, difficulty arises in so far as the matrix

[26] The Mahalanobis model, through possibilities of variation in the choice of λ_k, was also in principle capable of generating alternative time phased programs. The Mahalanobis model was further elaborated, by disaggregation with respect to intermediate goods (where the Mahalanobis model had been vertically integrated) and via the explicit assumption that capital goods producing the consumer (and intermediate) goods could not be shifted to producing capital goods (which *in any case* was implicit in the Mahalanobis model and indeed provided its basic economic rationale). This was done by Raj and Sen [140]. They produced a number of illustrative time paths whose value consisted in reemphasizing that an intertemporal choice had to be made, although no attempt was made by them to indicate how this might be done. Besides their model, while set in the context of an open economy, ignored the possibilities of trade offs between domestic production and imports for supplying any given bundle of goods at a point of time, as pointed out by Bhagwati [6]. For other comments, see Prasad [129].

Mention should also be made of a more novel, exploratory paper by P. N. Mathur [109] which gave a computational, time-phased solution to investment allocation decisions. Mathur, writing in 1962, explored the consequences of transforming an initial technological matrix for the Indian economy into technology of the U. S. type within a specified planning horizon. He used a linear programming formulation to discriminate between alternative transformation paths.

[27] See the article by S. Chakravarty and R. S. Eckaus [24].

$B^{-1}(I-A)$ may not be well behaved: thus there is no guarantee that the dynamic system which is governed by the matrix will ensure nonnegativity of the relevant variables over time. Hence, one cannot be sure that consistency necessarily implies viability. On the other hand, one could also work recursively backwards from an assumed terminal condition by using the finite difference version of Leontief's structural equations. The advantage of this procedure is that it would not require the B-matrix to be invertible. However, in all probability, we would fail to tally exactly with the initial conditions. If the magnitude of the difference between the historically given initial situation and the desired initial situation (thus worked out) was not large, then one could argue that the model provided some sort of a time-phased plan. But there is no a priori guarantee that the difference would not be significant; nor can we assume that the deviations would be found only in the sectors producing 'tradeables,' and hence be remediable by international trade.

Unlike the procedure discussed by Chakravarty and Eckaus, Manne and Bergsman [100] used a different method which gave what they called an "almost consistent model." In working out this almost consistent model, they did not rely on the complete solution of the non-homogeneous dynamic Leontief system. They computed a set of terminal output levels of $X(T)$ based on the need to reach specified levels of final consumption in the year T and to sustain growth in gross output at specified levels beyond the horizon T. This part of the exercise was therefore based on different assumptions from those underlying the determination of terminal investment levels used by the other models. Once $X(T)$ was determined, Bergsman and Manne obtained the timepath of $X(t)$ starting from $X(0)$, by log linear interpolation. Hence they had a unique planned output trajectory $X(t)$. As for the demand side, the timepath of final demand for year 0 to T was assumed to be given exogenously. But induced investment in both fixed capital and inventories was determined by planned output increases. Now, if we denote the planned requirements trajectories over time by $D(t)$, clearly $X(t) \neq D(t)$ for every t (and for every industry). Such differences were to be met by so-called "shock absorbers." These shock absorbers were either imports of producer's goods, or changes in projected consumption of food and fabrics and the domestic output of service sectors. Since construction, which was treated as a domestic service, turned out to be too severe a bottleneck for the initial year, they called their model an almost-consistent model.

The planning model first developed by Chakravarty, Eckaus, Lefeber and Parikh and later extended by Eckaus and Parikh [51], and hereafter described as the "CELP model," was formally the most detailed of all the models so far developed in the context of Indian planning.[28] This is not to suggest that it was completely adequate for the purpose of generating development plans for the Indian economy. But within the limitation of a linear model, the structure had sufficient flexibility to handle a number of important planning questions.

The model constructed by the above authors is best described as a finite horizon, linear optimization model involving explicit intersectoral and intertemporal relationships, which satisfies boundary conditions relating to the initial year as well as to the terminal year of the plan. This description indicates that the CELP model formally subsumed the structural features of the preceding models. In addition it provided an intertemporally optimal path of development which brought the economy from the initial situation to the

[28] This model has been described and discussed by Eckaus [49] and by Chakravarty and Lefeber [25].

desired terminal situation. It also distinguished between investment starts, investment in execution and completed investment.

The linear maximand used in the CELP model was the discounted sum of consumption over a five year period. The relative rate of discount over time was assumed to be constant to avoid 'regret' phenomena of the type discussed in the theoretical literature by Robert Strotz [167]. Consumption in each period was assumed to be of constant composition. In other words, there was no substitution allowed between the different items of consumption in one single period. This is an assumption of the Leontief variety on the side of consumption. These assumptions could be formally stated as follows:

$$(21) \qquad \text{Maximize } \sum_{t=1}^{T} W(t)C(t)$$

where $W(t)=(1+r)^{-t}$ and $[c]\ C(t) \leqq F(t)$ where $[c]$ is a diagonal matrix of proportional consumption coefficients with $\sum c_i = 1$. Clearly, r is the social rate of time discount whereas $F(t)$ is the vector of sectoral outputs designated for consumption. The authors expressed their unhappiness over their extremely rigid assumption with respect to consumption but justified their procedure on the ground that, for an economy such as India, the low level of per capita income did lend some credibility to the assumption of a fixed composition consumption basket.[29]

To make sure that consumption did not fluctuate from period to period, a linear model such as this required an explicit monotonicity constraint: $C(t+1) \geq C(t)$ $(1+n)$ where 'n' is a predetermined growth rate. Further, consumption in year 1, denoted by $C(1)$, was assumed to be greater than $\bar{C}(1)$, a predetermined amount. On the side of the structural relations, the CELP model specified interindustry relationships both on the current and on the capital account. The model differed from the traditional treatment on the side of capital formation by introducing a gestation lag of 3 years coupled with the assumption of an exogenous pattern of investment buildup. The model permitted the authors to assume that the pattern of buildup in investment was either uniform or different between sectors. These restrictions on the structure of the economy were described by the following set of relationships:

$$(22) \qquad \begin{aligned} AX(t) + F(t) + N(t) + H(t) + G(t) \\ + E(t) - M(t) - X(t) \leqq 0. \end{aligned}$$

$F(t)$, $N(t)$, $H(t)$, $G(t)$ were vectors of consumption, capital accumulation, inventory accumulation and governmental expenditure respectively. $E(t)$ was the vector of export levels and $M(t)$ was the vector of import levels.

With regard to capital formation the CELP model used the following set of relationships:

$$(23) \qquad bX(t) - K(t) \leq 0, \quad \text{where } [b] = \begin{bmatrix} b_1 & & \\ & b_2 & \\ & & b_n \end{bmatrix}$$

a diagonal matrix of sectoral capital-output ratios.

$$(24) \qquad \begin{aligned} K(t) - K(t-1) - Z(t) \\ + R(t-1) \leqq 0 \end{aligned}$$

$$(25) \qquad \begin{aligned} q^k Z(t) &= I^t(t-k); \\ \sum_k q^k &= 1, \qquad k = 1, 2, 3. \end{aligned}$$

$$(26) \qquad \begin{aligned} \sum_k p_{ij}^k I_j^{t+k}(t) - N_{ij}(t) &= 0; \\ \sum_j N_{ij}(t) &= N_i(t). \end{aligned}$$

[29] Clearly the assumption of a fixed-consumption pattern, in turn, implies the CELP model assumed identical, constant returns to scale tastes for each individual, thus ruling out explicit consideration of the question of the effects of changes in income distribution on consumption patterns.

Equation (23) states that total demand for fixed capital must be less than the capital stock currently available. Equation (24) states the balance relationship for net capital formation. $Z(t)$ is gross new capacity available in 't' and $R(t-1)$ shows the replacement requirements computed on any reasonable basis. Equations (25) and (26) describe certain structural aspects of the process of capacity creation. These consist of the assumption that additions to capacity consist of the blending of different sectoral outputs according to given proportions and at given moments of time. In other words, capacity additions are produced by outputs devoted to capital formation with a distributed lag structure. A well defined part of the intended capacity increase must be completed at $t-3$, $t-2$, $t-1$ periods in order to have the desired capacity increase available at period t.

If q^k denotes the proportion of total capacity increase that must be completed k periods in advance ($k=1, 2, 3$), then in order to have a unit of capacity increase in t, $I^t(t-k)$ must represent that part of $Z(t)$ that has to be completed in period $(t-k)$. This is shown by the equation (25)

$$q^k Z(t) = I^t(t-k); \quad \sum_k q^k = 1, \quad k = 1, 2, 3.$$

The lagged investment components $I^t(t-k)$ have their fixed coefficient production functions, one for each lagged period. In any one time period t, a given sector is going to contribute inputs for producing $I_j^{t+1}(t)$, $I_j^{t+2}(t)$ and $I_j^{t+3}(t)$. These inputs at time 't' are additive whether or not they are provided for capacity intended for $t+1$, $t+2$ or $t+3$. When summed, they make up the sector's contribution to gross investment in t. This is described in equation (26) where p_{ij}^k is the fixed production coefficient. Thus,

$$(27) \qquad \sum p_{ij}^k I_j^{t+k}(t) - N_{ij}(t) = 0$$

$$(28) \qquad \sum_j N_{ij}(t) = N_i(t).$$

In addition to capacity formation, there is also the equation for inventory accumulation, which is indicated by the following equation:

$$(29) \qquad H(t) = S[X(t+1) - X(t)]$$

where $S = [S_{ij}]$ is the diagonal matrix of inventory requirements.

Foreign trade problems were introduced in the model in a "complete" way, but not necessarily in a very satisfactory way. Export demand levels were assumed to be given exogenously. Hence no optimization was introduced there. Imports were divided into two categories: competitive imports and noncompetitive imports. Noncompetitive imports were related to sectoral production levels by fixed proportions. Competitive imports were related to the sectoral production levels through the device of import ceilings. Formally the competitive imports of the i–th type of commodity were given by the following inequality:

$$(30) \qquad \begin{aligned} &M_i^2(t) \\ &\leq m_i^2[FA(t) + \sum E_i(t) - \sum M_i^1(t)] \end{aligned}$$

where $M_i^2(t)$ was competitive import of the i–th type at time 't', $FA(t)$ was the foreign aid availability at 't', $E_i(t)$ was exports of the i–th commodity at time 't', $M_i^1(t)$ was noncompetitive import into the i–th sector, m_i^2 was an import ceiling. This meant that, after deducting from the total amount of foreign exchange earned at time (t) the total bill of noncompetitive imports, no more than m_i^2 times the residual could be allocated for competitive imports of the i–th type. This device of handling competitive imports through introducing boundary relationships was important in avoiding a pattern of complete specialization to which a linear model

would otherwise gravitate. Hence, despite the awkwardness of the procedure, it served an important purpose in view of the procedure of linear maximization adopted in the CELP model on the assumption that $\sum_i m_i^i > 1$.

The manner in which the investment figures for the terminal year were derived in the CELP model was different from that employed by Sandee, or Manne and Rudra.[30] The CELP model brought the post plan future into focus in terms of investment in the terminal year. This had some conceptual advantage over the converse procedure used by the earlier authors in as much as the growth rates were applied to the terminal year consumption component directly rather than to investment figures. This is readily seen as follows. From the purely mathematical point of view, the terminal configuration is merely the sum of the particular solutions corresponding to the nonhomogeneous elements of an open dynamic Leontief model. If we write the sectoral balance equations in the compact vector matrix form, as

$$(31) \quad X = AX + B\dot{X} + C + E + G - M.$$

where X is the vector of gross output level, C the vector of consumption, E the vector of exports, M the vector of imports and G the vector of government expenditure, then the sum of the particular solutions corresponding to C, E, M and G gives the terminal configuration. Investment is treated here completely endogenously, since we assume $I = B\dot{X}$. The complete solution to the nonhomogeneous part of the above differential equation is given by

$$(32) \quad \begin{aligned} X = {} & (I - A - rB)^{-1}Ce^{rt} \\ & + (I - A - \lambda B)^{-1}Ee^{\lambda t} \\ & + (I - A - \mu B)^{-1}Ge^{\mu t} \\ & - (I - A - vB)^{-1}Me^{vt}. \end{aligned}$$

[30] This procedure has been used independently by Manne. Also see Frisch [53].

Here r, λ, μ, v are the growth rates of C, E, G and M respectively.

We should note that since the CELP model assumed a fixed consumption basket, only the scale of consumption of the composite good was left to be determined by the logic of the optimizing mechanism. The items for the terminal year such as exports, government expenditure and imports were exogenously determined. Given these exogenous items, $K(T)$, which stands for the terminal vector of capital stocks, was expressed in this model as a function $f(r; T)$ where f is a vector function of the vector of the post Plan growth rates 'r' and the length of the horizon T; 'f' was determined implicitly by the solution of the optimizing mechanism.

Given the information on post terminal growth rates, the length of the planning horizon, the initial conditions, the structural equations, and the inequalities on the side of competitive imports, the model could work out a complete time path for all the variables such as production, consumption, and investment levels. (The model can, of course, be analyzed from the dual angle, e.g., in terms of the shadow prices of the relevant scarce factors. These shadow prices are the optimal rentals of different types of equipment as well as the price of foreign exchange. It is interesting to note that the shadow price of foreign exchange was always positive in this model since imports could always be used to increase the value of the maximand.)

The model as developed by the above authors was the subject of a critical appraisal by Srinivasan [165]. Many of Srinivasan's criticisms[31] were directed at the degree of weight to be placed on the numerical solutions thrown up by the model in some preliminary runs as well as

[31] For these comments, which are of considerable relevance to an evaluation of the Third Plan targets, Srinivasan [165] should be consulted. We turn to these questions shortly.

on the comparability of the solution to the model with the actual Third Five-Year Plan figures. Srinivasan, however, raised a *conceptual* question of general importance. This question related to the way in which the terminal conditions were specified in the CELP model. This model had envisaged terminal conditions as a way of sustaining post terminal rates of growth of consumption, where the composition of consumption was given exogenously but the scale was left to be determined by the optimizing mechanism. There are two limitations to this procedure. In a model involving an infinity of time, no indefinitely sustainable growth rate can exceed the growth rate of the labor force. Further, the composition of consumption, which in this model was pegged till eternity, could be expected to change if, as planned, income levels were going up year by year. Srinivasan, therefore, expressed a preference for setting terminal conditions in a way which would maximize indefinitely sustainable consumption per capita. This, of course, is none other than the disaggregated version of the so-called golden rule of accumulation. The terminal capacity vector in this case would be a function of the rate of growth of the labor force, and the production relationships of the system. Coupled with the assumption of full employment and a given time horizon, this would give the planner a vector of absolute levels of capacity needed at the end of the planning period. The method of setting terminal conditions along the lines suggested by Srinivasan has both conceptual and policy determination merits, especially in the context of a country like India with a massive and rapidly growing population.

Since the original form of the above model was published, moreover, Eckaus and Parikh [51] have done further work within the framework of the CELP model. While the conceptual structure of the model used by Eckaus and Parikh is essentially the same as that of the CELP model, the introduction of longer time horizons and also of new techniques in the agricultural sector by Eckaus and Parikh constitute important improvements.

Having stated the formal properties of the CELP model, and argued for its superiority over earlier efforts at operational planning models in the country, we now proceed to comment further on the analysis of the CELP model, indicating the areas in which further research is necessary and thereby highlighting some additional limitations of the CELP approach. We then conclude our survey of Indian planning models and techniques by discussing the precise manner in which the experiments with the CELP model were actually designed to throw light upon the important policy questions which Indian planners were concerned with at the time.

Possibilities of Further Improvement in Model-Making. The first important improvement in designing computable planning models would be to relax the assumption of linearity by introducing a nonlinear maximand. This is because linear maximization problems over time are known to display the so-called "flip-flop" behavior in consumption and investment levels, which is certainly very awkward for realistic planning models. If this "flip-flop" behavior is then sought to be corrected through the imposition of additional constraints, the constraints become more important than the optimization procedure, and hence the problem is not really solved. In the context of a multisector, intertemporal maximand, it is necessary to distinguish between two types of nonlinearity. First, there is the possible nonlinearity of a one-period utility function involving different types of consumer goods. Such a utility function would be an improvement upon the fixed coefficients approach. Secondly, the aggregate utility function over

time could be nonlinear bringing in considerations relating to the diminishing marginal utility of consumption as the consumption vector rises over time. Maximization of a nonlinear maximand however would raise many problems of a computational nature which may yet take time to solve but, with a Leontief type technology, the problem does not seem to be by any means hopeless.

Secondly, the foreign trade problem requires a more satisfactory treatment. Changing comparative advantage positions with respect to different commodities should be reflected more adequately in the structure of the model. On this point also, progress would require solving intricate questions of nonlinearity in somewhat the same way as in connection with the preceding objective.[32] However, in one respect, the nonlinearities relevant to an open economy would be somewhat more intractable to handle. These nonlinearities would arise when some learning phenomenon is operative or some other form of economies of scale is relevant, rendering the classical convexity property of the feasibility surface inapplicable. The programming problem in this context would then have a mixed integer form, for which no very suitable algorithm yet exists.

In addition to all these questions, we should also note that the structure of consumption and the techniques of production in the model are quite rigid and that this certainly restricts its empirical relevance. Further it may be valuable to introduce an explicit savings constraint, as political factors may impose a ceiling on the capacity to raise savings.

An additional feature which planning models devised specifically in the Indian context ought to take into account is the phenomenon of massive unemployment, disguised and open. No formal planning model developed in the Indian context *fully* takes into account the problem posed by such unemployment. The only way unemployment is reflected in these models is through the assumption that labor is free, thus enabling one to focus attention on scarce resources such as capital and foreign exchange. Since the clear implication of these models is that full employment can be reached only over a relatively long period of time, important questions arise regarding the way unemployment is to be handled in the intervening period. This not only raises problems of distributive justice, which are very considerable, but may also raise serious problems relating to economic efficiency if one remembers that, given suitable organizational efforts, unemployed labor may be put to creating extra social overhead capital, which would largely involve redistribution of aggregate consumption. We have noticed that this aspect of planning received considerable emphasis in the work of Vakil and Brahmanand which we discussed earlier. But we have also seen that the linkage between the deployment of labor in such labor intensive rural construction activities is not entirely independent of the decisions taken elsewhere in the economy with respect to the use of scarce resources including wage goods. This is because we do not have many cases in real life where labor can produce capital goods unassisted by capital goods, and further, there is no automatic mechanism through which potential savings of the agriculturists could be transferred into consumption of labor working on construction projects. Thus, there is always some linkage between labor used on social overhead projects and the pattern of industrial development that is envisaged. Thus, what is necessary is an

[32] On the problem of planning India's trade pattern within the context of linear one-period maximization models, one can refer to the thorough work done recently by T. E. Weisskopf. Weisskopf [175] does not consider the choice of export levels; instead he assumes these to be given.

explicit analysis of a planning model which would incorporate this important structural feature of the Indian economy and explore fully its implications. The models which have been developed so far, while they range considerably in their sophistication from simple to very elaborate constructs, have, however, been essentially concerned with the implications of the shortage of capital and foreign exchange, rather than with a full analysis of the abundance of one important factor of production, e.g. unskilled human labor.

Empirical Applications of the CELP Model. Despite these limitations, and essentially on an experimental basis, the CELP model was utilized to analyze certain important, policy questions.

Two essentially important problems were explored with the aid of the CELP model, which appeared contemporaneously with the execution of the Third Plan. (1) Suppose that the Third Plan terminal year (1965/66) capacity targets were accepted, and the initial conditions and structural coefficients also accepted as implicit in the Third Plan, was there a feasible timepath from the latter to the former? Further, aside from feasibility, was the implicit timephasing in the Third Plan "optimal" if the preference function involved a discounted sum of consumption over the five year period (1961/1966)? (2) Moreover, the CELP model was also used to generate an optimal timepath of investments and outputs, while replacing the exogenous terminal year capacity targets of the Third Plan and instead allowing endogenous determination of these targets via specification of the rates of post-terminal growth in different items of final demand. This latter trial run, therefore, raised the question of the optimality of the actual Third Plan in a more comprehensive manner.

The first set of questions was analyzed on the assumption that an annual aid flow of Rs.5000 million would be available through the Third Plan—an assumption shared by the actual Third Plan. Further, the targeted rate of growth of consumption was specified alternatively at $2\frac{1}{2}\%$ and 5% per annum; the rate of time discount was set at 10%. With these specifications and the structural coefficients assumed to be those implicit in the Third Plan, the actual Third Plan turned out to be feasible. However, it turned out that the *optimal* time-path solution for going from the initial to the terminal year capacity levels, under either assumption with respect to the growth of consumption,[33] required a lowering of consumption in the first year of the Third Plan as compared with the actual consumption that obtained in the year 1960/61 which preceded the Third Plan. Thus the crucial question on the feasibility side, if one were considering an optimal transition path from initial to terminal capacity levels (as defined in the Third Plan), was whether the planners were willing to be very austere in the beginning of the Third Plan. The optimal timephasing of the Third Plan, therefore, was not merely at variance with that implicit in the Plan itself[34] but, furthermore, it seemed unlikely to be feasible in practice.[35] Furthermore, quite aside from optimality, if the capital coefficients matrix used by CELP was indicative of the structural relationships in the Indian economy, then the investment figures in the Third Plan represented a serious underestimate.

The optimality exercise which allowed the terminal year targets to be endogenously determined, however (and which

[33] The difference between the two alternative growth rates with respect to consumption showed itself more in the derived figures for the first year consumption levels than in the maximum value achieved by the objective function over the planning period.

[34] Since the possibility of such a reduction in consumption in the first year was certainly not envisaged in the Plan document.

[35] It should be noted that varying the discount rate from 10% to 20% hardly changed the solution, especially with respect to the reduction in first year consumption.

therefore envisaged the possibility of an investment allocation different from that in the Third Plan itself), threw up a solution which was considerably different from the official Third Plan. First, the total discounted sum of consumption was considerably higher. The average annual rate of growth of consumption was considerably higher than the figure officially assumed for the Third Plan. Furthermore, the major constraining factors in this modified growth process turned out to be the consumption goods sectors rather than the sectors producing capital goods. Such a striking dissimilarity between the optimizing planning model and the official Third Plan was the subject of extensive comments by CELP [25] and by Srinivasan [165]. Clearly several important factors were responsible for this difference. One was the assumption of a fixed-composition consumption basket which was heavily weighted in favor of food, the chief consumer good. Secondly, agriculture was treated in the model as some sort of a "bargain" sector with highly favorable capital-output ratios and relatively small requirements for flow inputs of industrial raw materials. Furthermore, the length of the gestation lag of three years, which involved a substantial amount of investment in the pipe line so far as the capital goods sectors were concerned, together with a horizon length of five years, also affected the numerical results substantially. In the subsequent exercises done by Eckaus and Parikh [51], most of these assumptions were modified; the revised results showed a narrowing of the difference between the CELP results and the official Third Plan allocations, although the difference continued.[36] We might note here that the CELP results, pointing to a shift in investment allocations away from the capital goods sectors to the consumer goods sectors (primarily agriculture) became available at the same time as the new Indian Prime Minister, Lal Bahadur Shastri, was expressing views to the same effect and recommending publicly a shift in investments and policy attention to the agricultural sector. In view of the practical importance, therefore, of the issues raised by the CELP work, there ensued a lively debate. As it happened, the shift towards the agricultural sector was to come largely as a result of two successive droughts during 1965/66 and 1966/67,[37] which underlined the inadequate growth of agricultural output in relation to the growth of population and income, the consequent reliance of the economy on P. L. 480 imports and the necessity to push more systematically on the agricultural front. Further, the relative investments in the capital goods sector were to slacken off during this period (and until the moment of writing this Survey) owing to the significant decline in the availability of project aid for fresh projects and the slow completion of those in execution, and also the fact that the governmental investments were constrained by the shortage of wage goods owing to the two successive droughts.

Having discussed the general area of planning theory and techniques,[38] and analyzed the interplay between planning models and actual planning within the country, we now proceed to two related

[36] For further points on this difference, including the possibility that the official Third Plan might have had a different objective function implicit in its targeted allocations (even though no evidence of explicit and coherent argument on this question could be found in the Plan document) and that the Third Plan might after all make sense in terms thereof, see the original papers by Chakravarty and Lefeber [25] and Srinivasan [165].

[37] The resulting economic difficulties were to be accentuated by the Indo-Pakistani war of late 1965 and the concurrent stoppage of aid.

[38] By way of completion, we may also refer to the very recent work of Alan Manne and T. Weisskopf [102] which represents a dynamic, intersectoral model for India, 1967–1975, and is an extension of the CELP type research. Also, we should refer to two papers of interest, by A. Ghosh [58] and K. N. Raj [134], which survey the literature on Indian input-output research and planning models from different points of view.

questions that have led to extended discussion in India and have produced ideas of interest. These relate to (1) the choice of techniques and evaluation of investment projects, and (2) the issues of "financial" versus "physical" planning, "Keynesian" versus "quantity theory" approaches to planning without inflation, short term planning and spatial planning.

Choice of Techniques and Investment Criteria

The question of the choice of appropriate technology has been extensively discussed in India. The debate grew out of the preoccupation of Gandhians with the protection of traditional modes of production such as handweaving and home spinning in the cotton textiles industry (from elimination by competition in the market from modern techniques).[39] Much of the debate on this issue occurred concurrently with the formulation of the Second Plan. With its emphasis on the buildup of the capital goods sector, and also its concession to the demands for protecting the traditional forms of production, the Second Plan represented a curious blend of Soviet and Gandhian economic philosophies—consistent with the reputed Indian genius for reconciling the irreconcilables!

The economic discussion of the issue was largely centered around comparisons among three alternative techniques of weaving, representing different degrees of mechanization.[40] Raj [130, 131] whose pioneering discussion of this question was to prove fruitful, attempted to show that the rate of return on capital invested in any technique, if the capitalist rules of the game were followed, might not give a 'true' indication of the social rate of return on capital. He obviously had in mind the argument, earlier raised by Tinbergen [170] and others outside India, regarding the possible differences between 'accounting prices' and market prices in many underdeveloped economies. Raj also wanted to take into account the dole to be paid to the unemployed as a part of the cost of using the more mechanized technique, and to include the cost of social overhead involved in setting up highly mechanized new units, which could be largely avoided by sticking to more traditional techniques.

Raj's paper, which was published in the *Economic Weekly*,[41] generated an extensive debate in that journal, with both Indian and foreign economists participating therein.[42] The main thrust of the controversy that followed related to the question of how labor cost should be evaluated from a social point of view. The answer to this question turned, quite naturally, on how the social objective before the planner was defined. If, as Sen [153] noted, the objective was to maximize current output, "surplus" labor was "free" in terms of opportunity cost. If, however, the objective was to maximize the rate of growth of output, then the wage bill represented a

[39] Indeed, economic policy was to be deeply influenced by such political thinking, supported at times by economic argument. See Manmohan Singh [163] on this question, in the context of its impact on India's export performance in cotton textiles. As we will note later in the text, the Indian economic discussions ignored the effects on foreign trade, via quality effects, although these appear to have been empirically quite significant.

[40] We may also note here the interesting work of Dhar and Lydall [45] and Lakdawala and Sandesara [88] on the economics of small scale industries, which explored the problem of choice of technique, in this context, in a systematic manner.

[41] *The Economic Weekly*, now continuing as *The Economic and Political Weekly*, has occupied a unique place in Indian economic journalism. Founded and edited by Sachin Chaudhari, it has served as an outlet for tentative, economic ideas for many economists in the country, and with its semi-academic nature, it has helped to focus attention on many of the interesting problems facing the country. Among its contributors can be counted practically all the major economists, sociologists and political scientists in the country.

[42] Among the Indian economists in the debate were Ashok Rudra [149], Ajit Biswas [16, 17] and A. K. Sen [155]; the foreign economists were Joan Robinson [147], Charles Bettelheim [3] and Jan Tinbergen [171]. There were also several other contributors: [59].

social cost—provided we could assume that wages would be spent and would thus represent incremental consumption.[43]

However, the following comments on this position are warranted. (1) The conflict between current output and the rate of growth of output arises from the implicit assumption that the rate of savings is a function of the choice of technique and cannot be varied by fiscal policy to desired levels.[44] (2) Furthermore, from a policy point of view, we would have to assume that the government has adequate control machinery for fully regulating the choice of technique but not for varying the savings rate to any desired extent. (3) The argument that the wage bill represents the social cost of consumption must also be modified, as Raj had already noted, by the saving in consumption, if any, in the sector from which the labor "emerged." (4) Also, the formulation of the two objectives, maximization of current income and maximization of the rate of growth, involves juxtaposing two extreme forms of social preferences. The problem could instead be posed, following Ramsey, as one involving the evaluation of the entire streams of consumption associated with the choice of alternative techniques. This is, in fact, how Sen [156] proceeded later to look at the problem although his formal analysis emphasized the possibility of alternative timepaths and the necessity to choose therefrom, rather than formal optimization procedures.[45]

Furthermore, in the attempted application of these ideas to empirical problems such as the choice of technique in weaving, two major defects were evident. (1) The fact that alternative techniques may have different impacts on quality, and hence also on export performance, was not seriously considered. (2) Moreover, while computations of "reinvestable" surplus were made for each technique, it was forgotten that similar computations would have to be made all the way "backwards" to get a complete answer: the reinvestable surplus may be higher in technique A than in technique B if we take only the final stage of production into consideration, but the ranking may reverse if *both* direct and indirect reinvestable surpluses were taken into account.[46]

Finally, if we are to evaluate the entire debate as conducted at the time, it is indeed surprising that while the "true cost" of labor came in for a good deal of discussion, the associated concept of the "true cost" of capital did not figure much in the Indian discussion. It was, of course, mentioned from time to time that the market rate of interest did not reflect the true scarcity of 'capital' but it was not quite clear what was meant by such an expression. Some people argued that the current rate of return on capital invested by the companies for which balance sheet data were available constituted the index of scarcity of capital. But there were two difficulties in this case, one due to the unreliability of the estimate of capital as measured by the information given on bookvalues, and the other due to the difference that would normally exist between the average return and the marginal return. However, the question remains as to the relevance of the return on the capital even when one is referring to the marginal return.[47] Such a marginal rate of re-

[43] Little [93], writing independently later, formalized the arguments relating to the "real cost" of labor in the context of well defined models.

[44] This point was strikingly made in the early paper by F. Bator [2] on the subject of choice of techniques.

[45] Sen indicated however the resolution of this choice by maximizing the sum of consumption over a finite planning horizon: this would, however, amount to assuming implcitily a constant marginal utility of consumption. Subsequent attempts at dealing with this class of questions have recently been made by Dixit [46] and Marglin [104].

[46] The fallacy consisted in carrying over an argument, worked out at the macro level, to the evaluation of a micro industry.

[47] For an empirical attempt at estimating such a marginal return, see Eckaus and Lefeber [50].

turn would be an appropriate interest rate for discounting future benefits only insofar as we could assume that the capital market was perfect and hence society was equating the marginal rate of return on cost (in the Fisherian sense) with the marginal rate of substitution between consumption at two consecutive time points. In the absence of such assumptions being satisfied, the appropriate social rate of discount would diverge from the marginal rate of return, the degree of divergence depending upon what time profile of consumption is regarded as optimal by society.[48]

From the foregoing survey, it is clear that the Indian discussion of the choice of techniques was concerned primarily with the selection of an appropriate social rate of return on a unit of invested capital.[49] This problem came to be directly confronted eventually in connection with the cost/benefit analysis of the multi-purpose river valley projects, to which many economists turned their attention.

It would be tedious to attempt a detailed survey of all the studies; besides, their conceptual framework is largely identical. Hence we concentrate on three major analyses of investment projects, to illustrate the full range of methodological issues that were raised in the Indian context. These were the work of K. N. Raj [133] who dealt with some aspects of the Bhakra–Nangal project in the Panjab, of N. V. Sovani and N. Rath [164] who examined the economics of the Hirakud dam, and of Paul Rosenstein–Rodan [148] who analyzed the economic worthwhileness of nuclear prower production in India.[50] Rosenstein–Rodan's analysis differed from that of Raj and Sovani–Rath, in having *preceded*, rather than followed, the execution of a project.[51]

Raj's analysis was particularly acute in its discussion of the employment aspects of the project, and of its spillover effects in the form of indirect demands for consumption. Among the interesting conclusions he reached was that labor could *not* be drawn away in unlimited amounts at a given real wage rate: more labor could be had only by incurring higher costs to secure the migration of labor from more distant areas. Raj also found that the marginal propensities to consume of labor which migrated to the project site and of labor which came daily from rural households were different, the former having shifted to superior grains. He further noted several areas where the project design had ignored possibilities of substituting labor for capital, thus resulting in unduly high capital and import intensity. The design of the power project also suffered from not taking into account the time and the location pattern of the demand for power which was likely to arise: as a result, substantial investment in transmission lines, among other things, was made which could have been saved by building cheaper sources of power supply.

While, however, the analysis made by Raj clearly indicated that the time factor

[48] In fact, the social rate of discount, the "true cost" of labor and the optimal rate of consumption are all interdependent elements in a single, intertemporal optimizing problem. Hence, it would be inappropriate also to go about setting up numerical values for these magnitudes in an unconnected fashion, as is sometimes done. In this connection see the recent work of Dixit [46] and Marglin [104].

[49] The notion that the shadow foreign exchange rate may be higher than the parity also became more prevalent in the later stages of the discussion, although it had been obscured or omitted from many early analyses, both theoretical and empirical.

[50] Rosenstein–Rodan's analysis was essentially a critique of an earlier memorandum, advocating nuclear power production in India, by Homi J Bhabha, the late Director of the Indian Atomic Energy Commission.

[51] Indeed, one of the major deficiencies in Indian planning has been an inadequate appreciation of the need to analyze critically the economics of major investment projects prior to their approval and execution; there are indications, however, of forthcoming change in this respect. There is also considerable interest now among economists in conducting such studies of industrial plants and projects.

was crucial in evaluating an investment project like the Bhakra Nangal, which locked up an enormous amount of resources over a very long future, he did not make any attempt to compute the present-value estimates of benefits or costs associated with the project. Raj's reluctance on this point was understandable in view of our inability to say anything numerically very firm as to the social rate of discount which could be used to convert benefits and costs accruing at different points of time to commensurable units. But even the use of alternative sets of notional estimates would have been worthwhile to indicate the different margins of choice which existed in designing and implementing the project.

The analysis of Sovani and Rath corresponded to the usual type of cost/benefit analysis. Instead of computing an overall present value, however, they computed an annual cost/benefit ratio on the assumption of a given constant interest rate and a given depreciation flow every year. The interest rate was assumed to be 10%, on grounds which were not spelled out. The relative abundance of labor as a factor in the project design was not discussed; nor were the indirect effects of the project examined. Instead the authors concentrated on computing intensively what they considered to be the direct benefits associated with greater irrigated facilities and greater availability of hydroelectric power. While the use of the apparatus of cost/benefit analysis was a welcome feature of their exercise, the assumptions made with regard to the estimation of benefits and costs were somewhat drastic and could be questioned.

The substantive issues raised by Rosenstein–Rodan, by contrast with the exercise of Raj and Sovani–Rath, included the question of the suitable rate of discount. His calculations, based on data partly supplied to him by Ian Little, showed that the estimated unit cost of electricity generation was highly sensitive to the choice of the discount rate, because the capital costs of setting up nuclear power plants are extremely high. Further, Rosenstein–Rodan raised an interesting question with respect to the social, as distinct from the market, cost of inputs such as coal, which is an exhaustible asset with a highly uneven geographical distribution in India (implying significant transportation costs). Moreover, if a nuclear power plant were to be built, it would have necessitated heavy imports in the early stages. Thus the scarcity value of foreign exchange had to be guessed if the official exchange rate could not be regarded as an index of "true" scarcity.

Aside from all these considerations, there were uncertainties regarding technological progress, which might reduce the costs significantly in the future: the question of the optimal timing of the project was therefore equally important in this case and the possible postponability of the project was an additional factor to be considered. Even leaving out uncertainty, however, any answer relating to timing would require both an explicit assumption relating to the future demand and specification of the discount rate, even when the cost profile is known accurately: however, such an analysis was not undertaken by any of the participants in the debate.[52]

[52] In a nonstochastic context, this question of the optimal timing of a project has been analyzed by Marglin [103] [105]. He shows that the optimal time for construction occurs when the marginal loss in benefits from further postponement just equals the marginal savings in interest costs. Since this rule is too general, Marglin shows that under certain additional assumptions, the above rule could reduce to a simpler rule which says if we assume that the immediate benefit rate from a project will continue over its finite lifetime (whatever that may be), then the optimum time for construction is when the present value of benefits is not less than the cost of the project for the first time. Unfortunately, this simple rule, as Marglin is well aware, cannot be applied to situations involving increasing returns to scale. Since the argument for the nuclear power station was partly

Thus, while the discussion on investment projects which took place in India raised a number of important questions, no specific study of any investment projects undertaken so far can be said to have provided anything like a complete analysis in terms of the social cost–benefit calculus.

Other Issues Pertaining to Planning

Among the interesting debates in India, relating to planning, has been that associated with the notion of "physical" planning as contrasted with "financial" planning.

The phrase "physical" planning, attributed to Mahalanobis, was put forth as a counter to those who wished to opt for a smaller level of financial outlay on the ground that larger outlays would lead to inflation. The argument of the physical planners at the time was that if a set of investment targets was internally consistent and feasible in the sense that enough capacity and labor were available to produce it and the draft on foreign exchange did not exceed export earnings plus committed foreign aid, then there should not be any problem in generating enough "financial" resources to undertake the acts of investment.

But this contention, if taken at its face value, was spurious (despite the fact that it played a major role in getting policy-makers converted to larger outlays on investment). At its essence, the argument presupposes that if, "structurally" or "physically," we can raise the proportion of investment in national income, the necessary *ex ante* savings can be found to make the program "feasible" without inflationary effects. While this is a tautology at one level, its implication that policy-makers can therefore raise as much savings as are necessary to support the planned investments is clearly untenable.[53] There is no escape from having to investigate separately the problem of generating enough savings to support the projected investment, if inflation is to be avoided. Direct measures controlling the level and composition of consumption may be sometimes called for if the physical planning targets are to be implemented, a fact which was not fully emphasized by the 'physical planners'.

A related fallacy was the tendency, at the time, to argue that if India went in for capital goods production, it would not merely be *possible* to raise the savings–investment rate (by getting over the postulated transformation constraints) but that this would also *ensure* that the savings–investment rate would automatically be stepped up. The basis for this assertion was that "no one can consume capital goods so the only choice would be to have the higher savings–investment rate." It was forgotten that the effect could well be, for example, excess capacity in the capital goods sector or cutting into exports of (traditional) consumer goods (which could reduce imports of capital goods) *unless* fiscal policy was used to raise the savings rate to match the projected increment in the rate of investment.

While the notion of "physical" planning was used by the "big planners" (who favored larger investment outlays) against the more cautious "small planners," the actual fixation of the overall investment targets appears to have been more in the nature of a compromise between these

based on increasing returns through time, one could not therefore discuss its timing on the basis of the Marglin theorem. The resolution of the question requires an explicit intertemporal analysis of the particular investment projects in question, which was not undertaken by those engaged in this controversy.

[53] In a corn economy, for example, it may be possible to put all corn back into the ground and raise the ratio of investment to national income to 100%; but it would be patently absurd to argue then that, because this is technically or "physically" feasible, the necessary "financial" resources (i.e. savings) can therefore be found!

rival groups.[54] It has further been alleged by some critics that the savings–investment balance worked out in the Plans suffered from an inherent inflationary bias even when the numerical projections looked quite safe on paper. These biases arose from (1) exaggeration of productivity estimates (especially in agriculture), (2) liberal guesses at the 'safe' limits to deficit financing, (3) underestimation of investments in private agriculture and small industry (which generally were 'cut' by the big planners to reduce the total investments, in full knowledge that no reduction would *actually* follow as these were mere estimates and fairly beyond policy control), and (4) failure to make allowance for additions to inventories, not to mention (5) the 'uncovered gap' in the resource balance exercise which was actually published in the Second Plan document.[55]

In this context of fiscal policy programing, which examines whether *ex ante* savings would match the projected investments, we may note that the "Keynesian" approach has now come to stay in India. It has become customary for the Plan documents to present fiscal policy in the context of an overall savings and investment exercise, rather than purely in budgetary terms. This has been the case despite a tendency toward thinking in quantity–theoretic terms on the part of some economists in India.[56]

The question of fiscal policy also raises the entire question of ***short–term*** (or annual) ***planning*** in India. There is little doubt that short–term planning, which takes a narrower horizon than the Five Year Plans, is essential in a country such as India where (i) agriculture plays an important role (in exports, industrial production and consumption) but is subject to wide fluctuations in performance—as was brought home to Indians during the two successive and serious droughts during 1965/66 and 1966/67; and (ii) foreign aid also is crucial (in project investments and utilization of industrial capacity). The need to adapt planned expenditures, and economic policies in general, in the light of fundamental revisions relating to critical assumptions, is certainly obvious.[57]

Unfortunately, there has been no serious attempt so far at building a relevant short–term model for the Indian economy, which would permit the policymakers to take informed decisions—a lacuna which became only too obvious when the recent

[54] The cautious groups have belonged to the Finance Ministry whereas the big planners have been associated with the Perspective Planning Division of the Planning Commission.

[55] Pointing to these factors which build into the Plans a systematic tendency towards price inflation, despite the apparent balancing of *ex ante* savings and investment, Bhagwati [10] has observed that such "over-extended" planning is likely to have led to too many starts and too few completions of projects, thereby leading to a serious decline in the overall level of productivity. He has also argued that the protagonists of such a strategy may have thought that, "getting exaggerated Plan targets accepted by the Government would push it into the extra effort (via domestic taxation) which otherwise would not be forthcoming. This strategy presumably relied on the built–in creation of a mildly inflationary situation to put pressure on the Finance Ministry to tax even more than its Plan commitments, quite ignoring the fact that the resulting unrest in the urban areas could possibly prevent the Finance Ministry from taxing even as much as they would have (if the inflation had *not* been deliberately built in) and thus reducing the resultant level of real investments as also their efficiency."

[56] On this issue, see Little [92], Chakravarty [21] and Padma Desai [43]. Further, Little's [94] systematic work on taxation for the Third Plan, for the PPD, was one of the first attempts at conducting the overall savings–investment exercise in some detail, although the Second Plan did contain the rudiments of essentially the same type of exercise. Coming from the side of the big planners, Little's work appears to have had an influential role in the favorable outcome for the big planners.

[57] In this connection, it is interesting that there has been no extended discussion of suggestions such as the "rolling plans" discussed in the Swedish literature. However, the question *has* led to demands such as, for example, having a "core" of "basic" projects and expenditures which would be carried through in any case, with more being done if aid comes through. However, there has been no systematic exploring of these issues.

industrial recession, in 1967, left the policymakers in a position where action had to be taken without any systematic knowledge of key relationships (such as the effect of corporate taxes on savings, pricing policies of corporations etc.).[58] We should note here, however, an interesting, early attempt towards building an econometric model for India, by Narasimhan [122]. Narasimhan constructed an econometric model roughly on the same lines as Tinbergen's pioneering model for the United States and the United Kingdom. Narasimhan's model had 18 equations, of which 7 were definitional, one was institutional and the remaining were behavior equations. While Narasimhan's was a pioneering attempt, he appears to have paid insufficient attention to adapting the Tinbergenesque models to the very different structural features of the Indian economy. These would relate, for example, to the distinction between urban and rural propensities to save, the elasticity of the marketable surplus in agriculture, the distinction between factory employed labor and self employment, and so on. In view of these limitations, therefore, Narasimhan's exercise is best regarded as an exploration of the question of building a suitable, short–term econometric model for India, rather than as a realistic and usable model.[59]

Finally, it is interesting to note that, while Indian planning models have become fairly sophisticated in relation to intertemporal phasing and perspective planning, there has been no comparable extension of analysis to questions of *spatial* planning. This is somewhat surprising in a country with a federal setup and where the constituent States have come to follow increasingly inward looking policies.[60] An important consequence of the lack of spatial planning of industrial targets has been the tendency for the targeted industrial capacities in each industry to be competed for by numerous claimant States, thus resulting in the allocation of plants with uneconomic scale to as many States as politically necessary.[61] While it would certainly be naive to expect that efficient allocation of industrial targets among different States, consistent with the satisfaction of constraints with respect to *aggregate* levels of industrial investments in each State on political grounds, would necessarily be accepted as a politically satisfactory method of spatial planning, it is nonetheless true that few economists gave serious attention in India to this question. An important, recent departure in this respect is the work of Srinivasan and Manne [99] which has brought original analysis to bear upon the policy question of optimal location, size

[58] We may record, however, that in this area as well, more systematic work is beginning to emerge. An example is the detailed study of inventory holdings in the large scale, manufacutring sector by Krishnamurty and Shastri [85] which illuminates an otherwise obscure area of the Indian economy.

[59] In this connection, we may also refer to a short-term policy model built by Padma Desai [41] for India, which was of an illustrative nature. Her model was based on a (3×3) input–output table which was closed with respect to consumption. The exercise was interesting because of the way in which she determined consumption endogenously via the assumption that the propensities to consume of workers in different industries were different. The model could thus predict the effect of changes in exogenous items of final use on sectoral output levels more completely than the input-output models discussed earlier in the text. However, the model did not specify how the instrument variables at the command of the government were related to the exogenous variables in the input–output sense, so that it was not sufficient for generating a shortterm plan in a complete way.

[60] We take up this question again in the next section, when we discuss the food zones that have operated in India.

[61] This process has been noted, and the lacuna in Indian planning techniques criticized by several economists: see Bhagwati [9] and Hanson [62], in particular. Bhagwati [9] has also emphasized that the uneconomic scale plants may well be economically justified if the spatial distribution of demand and transport costs, for example, make centralized production uneconomic, and that it is the lack of economic analysis of this issue, rather than the actual solutions implemented, which is unsatisfactory.

and expansion of industrial capacities.[62]

Having then discussed the principal issues and models which have emerged in relation to Indian planning, we now proceed to the Indian analyses that bear on questions of agricultural policy.

II. *Agriculture*

Indian agricultural policy discussion has taken place against the background of a trend rise in agricultural production, especially of foodgrains, which has fallen sufficiently short of the growth in demand arising from income and population growth to require continual and significant import of foodgrains under the P.L. 480 program. In consequence, economic analysis has largely been concerned with questions relating to agricultural price and distribution policy, and also the economic efficiency of alternative forms of land tenure and agrarian organization.[63]

In turn, these questions have led to analytical work on a whole range of problems with a direct bearing upon policy decisions.[64] Prominent among these studies have been the analysis of (1) the economic "rationality" of farmers, (2) the response of marketed surplus and production to price changes, (3) the relationship of land tenure systems and agrarian organization to the efficiency of factor use and to the elasticity of marketed surplus, production and investment to price change, and (4) the question of the existence and measurement of disguised unemployment.

Furthermore, Indian economists have also turned increasingly to efficiency questions relating to public agricultural investments. Economic analysis has been increasingly deployed, principally by Minhas [114] and Minhas–Srinivasan [115] to examine problems such as the efficient allocation of irrigation water and fertilizers,[65] although the choice between alternative ways in which farm output may be raised (e.g. land reclamation *versus* intensive cultivation) has not yet been fruitfully explored at an empirical level.

Agricultural Performance

At the outset, we may note that the production performance of Indian agriculture has been the subject of lively debate.[66] Pointing to India's continued reliance on P.L. 480 imports, economists such as Dandekar [33] have tended to dismiss India's agricultural performance as dismal. On the other hand, Raj [135] and Dantwala [34], while conceding the inadequacy of this performance, have attempted to put it into perspective by noting that the annual, compound rate of growth of production of foodgrains at 2.98 per cent and all commodities at 3.19 per

[62] For general work on "regional" models in India, the survey by Ghosh [58], which we have referred to earlier, is a valuable reference. Aside from the many references there, work on transportation and regional planning models has been done in India by several other economists, including M. Datta-Chaudhuri [39] and K. Sundaram [168].

[63] With respect to agrarian organization, the *social* aspects of alternative policies have also claimed equal attention in the Indian discussions. Cf. Dantwala [35]: "It may be pertinent to enquire as to what has provided the main inspiration for the proposal to impose a ceiling on individual ownership of land: the urge for distributive justice or the necessity of a more rational use of the land surface? The impromptu answer would perhaps be: both. But it would be honest to admit that the prime motivation is distributive justice. In the context of the acute land hunger and millions of dwarf farms, ownership of large areas of cultivated land by a few is considered highly inequitable, justifying the imposition of an upper limit to individual ownership."

[64] *The Indian Journal of Agricultural Economics*, currently in its 22nd volume, is an excellent guide to the full range of problems that Indian agricultural economists have considered from time to time. As already noted earlier, our survey is necessarily selective.

[65] We may also recall here the cost–benefit analysis of Sovani and Rath [164] and Raj [133] which we have surveyed in Part I. See also the excellent review of the literature on application of economic theory to Indian agricultural policy discussion by Khusro [76].

[66] We may note here the important work of Minhas and Vaidyanathan [116] in measuring the rate of growth of Indian agriculture, by 268 districts, for the aggregate output of 28 major crops for 1951–54 to 1958–61. For their method of measurement, including their decomposition of this growth into crop pattern change, productivity and acreage "effects," as also a valuable survey of other work in this area, see Minhas [113]. Parikh's work [127] in this area is also noteworthy.

cent for 1949–50 to 1964–65 does not compare unfavorably with performances in most other countries including those in South East Asia.[67] Moreover, Dantwala [34] has also noted that, contrary to general belief, productivity has also increased through this period, with acreage under all commodities increasing by 8 per cent but production by 34.8 per cent.[68]

Whether one regards the agricultural performance, however, as dismal or just inadequate for India's developmental needs, the pertinent questions are whether (1) governmental policies (especially concerning prices) could have improved it, and (2) governmental policies (especially with respect to internal procurement, imports, private trade and public distribution), were efficient, *given* the agricultural performance. Before we discuss these two principal policy questions at some length, we survey the Indian analysis of the empirical relationships, pertaining chiefly to the marketed surplus and production of foodgrains and overall agricultural production, that have a direct relevance to these questions and have indeed entered into the controversies surrounding these questions.

Economic Rationality in Agriculture

We begin with the literature on the "economic rationality" of the agricultural sector. Whether agricultural and rural people and institutions respond to economic motivation or are impervious to it is a general question, of which the possible response of marketed surplus and production to price changes (which we discuss in the ensuing sections) are only the most obvious examples. The Indian literature on the broader question of economic rationality in this sector divides itself into empirical analysis aimed at (1) examining the efficiency of factor use within the existing institutional framework, and (2) demonstrating that the institutional framework itself adapts to the profit motive.

(1) Among the principal contributions to the former class of questions is Hopper's [70] analysis of the efficiency of Indian farmers in the allocation of resources. Hopper's method is to estimate production functions for his selected crops[69] and demonstrate that the factors used indeed earn the value of their estimated marginal products. However, as Nowshirwani [123] has correctly pointed out, Hopper's single equation estimation of his production functions leads to estimates that are neither unbiased nor consistent (in a statistical sense) so that Hopper's results must be treated with some scepticism.

Furthermore, we must note a different approach by D. K. Desai [40] to the problem of efficiency of factor use, which aims at discovering the optimal utilization of *existing* resources on individual farms and contrasting the resulting utilization pattern and returns with the actuals. Desai uses linear programing methods for this purpose, utilizing data collected for the Farm Management Studies in two districts of Maharashtra during 1954–55 to 1956–57 and examining forty *individual* farms. This pioneering work has reached the conclusion, at variance with Hopper's, that there is often a significant gap between possible and actual returns to farming, indicating economic inefficiency. However, as Hanu-

[67] Dandekar [33] has noted that if the two (drought) years 1965–66 and 1966–67 are included, the performance looks even less satisfactory; however, Dantwala [37] has pointedly replied that the 1967–68 crop, which is at a bumper level, would bear out *his* notion of the trend. In this context, we may also note that Raj's [138] comparison of the agricultural growth rates in India and Pakistan is a useful corrective to Mason's [106] adverse, comparative view of India's agricultural performance.

[68] However, the performance through the entire period conceals a serious deceleration which sets into the overall growth performance as also in the growth of average productivity, with the Third Plan. Hence, the agricultural performance may have been not merely inadequate but also steadily becoming worse. The 1967–68 crop, however, has been a bumper crop.

[69] The data relate to one Indian village and to the expected output from invested resources of its farmers in a single agricultural season.

mantha Rao has pointed out to us, the gross inefficiencies which Desai's analysis indicates are probably to be attributed to the fact that the results are derived by references to single period, *ex post* price vectors. If *expected*, rather than *ex post*, prices were considered, the results might be significantly different. A similar doubt attaches to Desai's use of given production relationships, which again may lead to inefficiency in the design of the test for optimality: expectations with respect to weather, for example, may significantly alter the crop pattern that may be adopted on efficiency grounds. Apropos of this discussion, it is also pertinent to remember the important distinction that Lipton [91] has drawn between farmers' production response to price change (which we discuss later) and profit maximization. For example, producers responding to prices in a cobweb model may not be maximizing profits in the long run.[70]

(2) The analyses which seek to establish that the rural institutions change in response to economic motivation are of equal interest. While a sociologist such as Scarlett Epstein [52] has attempted cross-section analysis to show this, by contrasting two villages which differ only in terms of the recent availability of irrigated water, Raj [139] has noted that the "sacred cow" is not so sacred after all and manages to get slaughtered even in Hindu-intensive areas if ecological and price factors make it economically advantageous to do so.[71]

Hanumantha Rao's [67] analysis of share cropping is also of considerable importance in this context. Rao shows in an ingenious fashion that share cropping obtains generally in those areas, and for those crops (such as rice and wheat), where the element of innovative management and entrepreneurship is minimized because of lack of significant substitution possibilities among rival crops and factors, and the element of *uncertainty* is thus reduced to negligible levels: "crop-sharing arrangements are extensive under relative economic certainty and fixed contractual payments where the degree of uncertainty is high."

A few comments on this novel idea are in order. (1) Rao's hypothesis, which seems consistent with the cross-sectional facts of the Indian situation as of any one period, would lead to the further refutable hypothesis that, as technological possibilities for application of new inputs such as better seeds and fertilizers are introduced, share cropping would give way to other forms of tenurial relationships. With the introduction of these new techniques in India during the last few years, such an empirical test should be feasible.[72] (2) Further, even if Rao's explanation of the "rationality" (in the sense of its consistency with profit maximization) of share cropping is valid, how are we to interpret its effect on efficiency of the utilization of available resources? An alternative formulation of the rationality of share cropping, in the spirit of Rao's argument, provides a clue to the optimality of share cropping as an institution. If we focus on the *stability* (or stationariness) of the agricultural techniques and acreage allocation possibilities, it is possible to argue that share cropping is consistent with optimality because the shares would come to approximate the level where factors tend to earn the value of their (stable, long-run) marginal product. Share cropping would then yield, on the average, the same results as capitalist

[70] Lipton [91] provocatively entitles his review of Dharm Narain's work: "Should Reasonable Farmers Respond to Price Changes?".

[71] Raj's cross-section investigation of this question is of further interest because he considers the important, related question of how the cattle stock should be evaluated from an economic point of view.

[72] Rao explicitly notes that: "Crop-sharing may cease to be a beneficial arrangement as modern profitable inputs assume significance. The incentives for increased investments as well as for capturing the returns on such investment may lead to the preference for fixed contractual payments."

methods in a situation of long-run stability characterized by stagnant technological possibilities. A possible test of this hypothesis would be to examine share cropped farms, over a period characterized by such stability, for efficiency of factor-use in the sense in which Hopper [70] has done for his village.[73] Clearly, in any case, Rao's analysis opens up a fruitful area for further empirical investigation.

Behavior of Marketed Surplus

Indian economic analysis has concerned itself with two principal questions relating to the marketed surplus of agricultural foodgrains: (1) does this surplus vary directly with the relative price of these goods or does it behave "perversely"; and (2) what is the share of holdings of different sizes in the supply of marketed foodgrains? The former question has direct and obvious relevance to the issue of agricultural price and tax policy whereas the latter bears on the important question of the economic effects of land reform which involves regrouping of landholdings into different sizes, whether towards smaller holdings via measures such as landholdings ceilings or towards larger holdings via measures such as legislation preventing further fragmentation. Further, the two questions have, in turn, been linked in the analytical discussion by economists who argue that the price response of the small and large holdings with respect to marketed surplus is not similar.

Response to Price Change: Essentially, the question at issue, in analyzing the response of marketed surplus to price change, relates to the elasticity of the Marshallian offer curve of the sector supplying the surplus. And this is indeed how Raj Krishna [80] and T. N. Krishnan [86] have explicitly formulated their analysis of this question. Furthermore such a formulation of the problem directly indicates that positive and negative price responses are both "normal": the offer curve may readily have a backward bending stretch where the elasticity of supply of the (foodgrain) surplus with respect to price change is negative.

Nonetheless, the Indian debate on this issue has elicited arguments, mainly by Khatkhate [71] and Khusro [75], which aim at establishing a priori the "normalcy" or the inevitability of a positive or a negative elasticity. More interesting, however, have been the analyses, principally by Mathur and Ezekiel [110], which have attempted to explore the issue a prori within a framework which differentiates between different size-classes of holdings. At the same time, Raj Krishna [80] and Krishnan [86] have made perhaps the only systematic attempts at indicating what the likely elasticity of the offer curve for foodgrains might be in the Indian context.

Khusro's [75] analysis reaches the "strong" result that farmers will retain more and hence market less, out of given foodgrain production, if the market price is lowered. This is the consequence of his method of analysis which implicitly puts restrictions on the shape of the offer curve. Taking a box diagram as in Figure 1, where AZ is the given output of foodgrains, Khusro draws in PR' as the curve showing diminishing marginal utility of produce retained for consumption and SS' as the schedule representing diminishing marginal utility of sales. Then, equilibrium is at D where the amount of marketed surplus is EZ. With a lower price for foodgrains, Khusro shifts the curve SS' down to ss', when G becomes the new equilibrium point and the marketed surplus has decreased to FZ. This conclusion, however, is the result of two highly restrictive assumptions: (i) separable (and hence cardinal) utility, explicitly noted by Khusro; and (ii) sufficient restrictions on the rate at

[73] In this connection, it is suggestive that Hanumantha Rao's comparison of the share cropping rental with the output per acre on "capitalist" farms [67, Table 6] shows a similar intensity of cultivation, per acre, on both sets of farms.

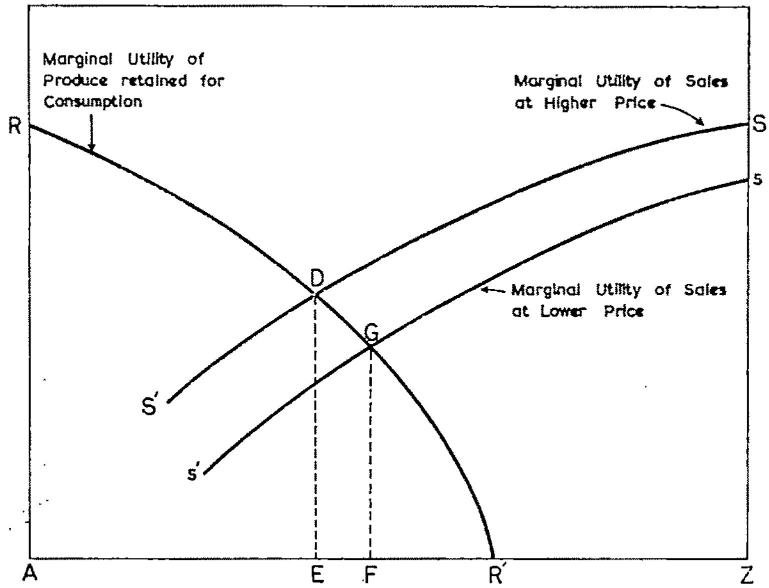


FIGURE 1

which the marginal utility of nonfoodgrains falls vis–a–vis the price difference in the two situations, to ensure that the SS' and ss' curves do not intersect.[74] If either of these assumptions is relaxed, the possibility of a negative elasticity of marketed surplus will reemerge.[75]

As with Khusro's attempt, other economists have attempted to argue exactly the opposite proposition on a priori grounds. Thus, for example, Khatkhate [71] has argued that the (small scale) Indian farmer will increase his marketed surplus when price falls "in order to maintain the same level of money income." Mathur and Ezekiel [110] have similarly argued that the subsistence farmers have an inelastic demand for cash and hence "if prices rise, the sale of a smaller amount of foodgrains provides the necessary cash and vice versa. Thus prices and marketable surplus tend to move in opposite directions." The fixed cash needs which these authors have in mind are debt obligations, rent, land revenune and a Ricardian-type bundle of nonagricultural subsistence goods. However, this argument seems tenuous. It implies that, at subsistence level, there is zero income elasticity of demand for commodities other than the foodgrains (produced on the farm) and also a zero substitution effect: neither assumption would appear to be logically or empirically inevitable at a "subsistence" level of farming and income, no matter how subsistence is defined.[76]

Dandekar [31], in examining the Mathur–Ezekiel thesis, has argued that a

[74] Khusro [75, page 278] outlines the second condition as well, though not quite fully. While he mentions that the rate of diminishing marginal utility of sales receipts should be "small," he omits reference to the fact that the price change may be "large" and may outweigh the "smallness" of the diminishing marginal utility, resulting in the interesection of the two schedules and the reversal of his proposition.

[75] Of course, the entire argument relates only to the sign of the consumption effect as price changes. If production were also allowed to vary in Khusro's exercise, and if it was positively price elastic, then the price elasticity of marketed surplus could well be positive despite the consumption effect being positive.

[76] Also note that the Mathur–Ezekiel assumption of "fixed cash needs" is *overly* sufficient for deriving a negative elasticity of marketed surplus. Further, as Nowshirvani [125] has noted, the negative unitary price elasticity for marketed surplus is implausible on dynamic stability grounds, since it would require the urban elasticity of food demand to exceed unity, which is an implausible assumption.

negative price elasticity of marketed surplus would characterize the farms which are not too close to subsistence and hence are large enough to supply marketed surplus and therewith earn cash income and which are at the same time not large and prosperous enough to show a "normal," positive price elasticity of marketed surplus. Dandekar argues that, for this size–class of holdings, the negative price–elasticity is readily explained, not by reference to a "fixed cash needs" hypothesis, but "as normal consumer behavior in the face of changing income. When the price of the crop which they produce changes relative to other prices, the real income of the farmers in effect changes. With lower relative prices for what they produce, their real incomes are in effect lower and as consumers they are worse off. Under the circumstances, they behave like other consumers at their income level would do. They must consume a little less of everything, food and non–food alike even if they happen to be producers of food. This is what they do. They consume a little less of their own produce and consequently sell more of it on the market." Dandekar reaches this apparently obvious conclusion by either ignoring the substitution effect or by implicitly assuming that the income effect will dominate the outcome.

Dandekar's analysis of the elasticity of price response by other size–classes of farmers is also not persuasive. With respect to the farmers at the bottom end of the scale, he argues that the absence of marketed surplus of foodgrains and the dependence "mainly on other means such as sale of other crops or wages earned from farm and off–farm employment or even remittances received from family members working in cities" implies that the question of the elasticity of marketed surplus does not apply to them. However, this argument rules out the possibility of a marketed surplus arising in response to price rise, for example, via shift in resource allocation toward foodgrains and/or curtailment in own consumption if the substitution effect is large. Dandekar's further contention that, in many cases, the farmers in this size–class of landholdings actually buy foodgrains on a *net* basis, is also inconclusive in this respect: these farmers can still turn into net suppliers if the price change is favorable enough and the consumption effect has a negative sign. Moreover, even on an empirical level, Dandekar's citation of the data for sale and purchase of Jawar by the small landholdings in the Akola and Amraoti districts of Madhya Pradesh for 1955–56 is not conclusive. The fact that these farmers purchase *jawar* on a net basis does not necessarily rule out their selling *other* foodgrains (such as wheat, which is also produced and sold in these districts), so that the *net* position on *overall* foodgrains supply may be different from that on the supply of only *jawar*. More systematic empirical analysis of this question is clearly necessary.

Similarly, it is hasty to conclude that the "large" holdings will necessarily have a positive elasticity of marketed surplus, thanks to a "well behaved," negative consumption effect. It is easy enough to imagine, for example, a lower price of wheat leading to greater supply of marketed surplus thereof as the farmer shifts to increased consumption of the inferior cereal, *jawar*.

Despite these qualifications, it is clear that empirical analysis of the price response of marketed surplus, which differentiates between different size–classes of landholdings, is likely to be more insightful than an aggregative analysis.[77] Unfortunately, however, the only serious attempts at empirical analysis of the

[77] Since foodgrains also represent a complex of different cereals, the analysis would have to contend also with (1) their price substitution in production and consumption; and (2) the possibility that some of them may

problem in India have been at an aggregative level.

Using Rural Credit Survey data on the market surplus of foodgrains, Krishnan [86] has estimated the constant elasticity demand function: $r\bar{Q}=AP^{-\alpha}(\bar{Q}P)^{\beta}$ where $\bar{Q}$ is the total output of foodgrains in the short-run, P is the price of foodgrains, $\bar{Q}P$ is the income of the farmers and r is the proportion of output consumed by the farmers themselves. He estimates the elasticity of marketed surplus to be -0.3030 for the period 1959–60 to 1962–63. Note, however, that if the price elasticity of production response, which has been estimated to be positive in other studies surveyed elsewhere by us, were to be admitted into this exercise, and the assumption of a constant output thus relaxed, the price elasticity of marketed surplus could well become positive.

Raj Krishna [80], in fact, has experimented with different ranges of price elasticities of output and consumption to establish that, in the Indian context, the price elasticity of marketed surplus of a *single* subsistence crop (as distinct from Krishnan's estimate for all foodgrains) is indeed likely to be positive. Starting with the simple identity:

$$\frac{dM}{dP} \equiv \frac{dQ}{dP} - \frac{dC}{dP} \tag{33}$$

where M is the marketed surplus, Q is production, C is self-consumption and P is the price, Raj Krishna further decomposes the consumption term into income and substitution effects and then proceeds to put hypothetical ranges of values on each of the parameters in the derived expression for the elasticity of marketed surplus.[78] Unfortunately, however, his decomposition of the consumption term omits the income effect which follows from the change in the value of the initial consumption as price changes. Nowshirvani [124], who spotted this error, has reworked Raj Krishna's analysis and, on using the same empirical ranges of values for the parameters relating to production and consumption response, finds that it is not possible to rule out negativity in the price elasticity of the marketed supply of foodgrains in the Indian context.[79]

Distribution of Marketed Surplus by Size-Classes of Landholdings: Indian analysis has also extended to the question of the shares of different size-classes of holdings in the marketed surplus of foodgrains and agricultural produce in general. Unlike with the question of the price elasticity of marketed surplus, however, this problem has attracted mainly *empirical* analysis. Among the contributions to this area of research, Mathur's [107] findings on the marketed output of *jawar* in Akola and Amraoti for 1955–56 and Dharm Narain's [120] patient compilation and processing of the relevant information for the marketed surplus of *agricultural produce* for India in 1950–51 are of interest. Raj Krishna's [82] attempt at cross-section analysis of the marketed surplus, by size of farm *output*, for rice and wheat for selected markets has also attracted considerable controversy in this context.

While these studies agree on the proposition that the supply of marketed surplus is not a characteristic of only the "large" landholdings, they differ on several other points of substance. While Dharm Narain finds, for example, that even landholdings in the size-class 0–5 acres contribute as much as 20.7% of the value of their output as marketed surplus (Table 1), Mathur's study shows that the sale of

be inferior goods in consumption. In consequence, a sharp distinction would also have to be drawn between the price elasticity of the marketed surplus of a single foodgrain and more.

[78] These values are not entirely hypothetical but are derived by him from several studies, including his own study of the production response to price change in the Panjab [81].

[79] We may emphasize however that the negativity of price elasticity does not imply the unitary elasticity assumption of Mathur and Ezekiel.

TABLE 1—DISTRIBUTION OF MARKETED SURPLUS BY SIZE-GROUPS OF HOLDING

	I			II		
Size of holding (acres)	Marketed Surplus (Rs. crores)	(1) as % of value of output	(1) as % of Total Marketed Surplus	Marketed Surplus (Rs. crores)	(4) as % of value of output	(4) as % of Total Marketed Surplus
	(1)	(2)	(3)	(4)	(5)	(6)
0–5	266.7	20.7	24.9	564.0	33.6	26.0
5–10	175.8	14.1	16.4	444.8	27.4	20.5
10–15	54.7	9.7	5.1	170.1	23.1	7.9
15–20	80.1	18.2	7.5	172.8	30.1	8.0
20–25	54.0	20.4	5.0	111.0	32.2	5.1
25–30	65.4	28.8	6.1	116.8	39.7	5.4
30–40	80.5	29.9	7.5	139.6	39.8	6.4
40–50	67.8	38.0	6.3	107.8	46.4	5.0
50 and above	228.0	44.8	21.2	339.9	51.4	15.7
Total	1073.0	21.5		2166.8	33.4	

Source: Dharm Narain [120, 35].

jawar by landholdings in the size-class 0–5 acres is practically nil. While this discrepancy in the results may arise from differences in the concept of the marketed surplus and also in the period and area covered, it arises undoubtedly also from the fact that Mathur's findings refer to a single crop, *jawar*, whereas Dharm Narain refers to agricultural produce in general. As Dandekar [31] has noted, the All-India Rural Credit Survey has shown that this size-class of farmers are typically net purchasers of *foodgrains*, while earning cash income by "*sale of other crops* or wages earned from farm and off-farm employment or even remittances received from family members working in cities."[80]

Another difference of considerably greater policy relevance relates to Dharm Narain's finding that the proportion of output marketed falls until the size-class 10–15 acres is reached and rises rapidly thereafter. An important consequence of this finding, if statistically valid despite the numerous "adjustments" that have gone into its derivation, would be that land ceilings resulting in breakup of the larger holdings could increase, rather than diminish, the marketed surplus of agricultural produce—ignoring, of course, the derivative effect of any resulting shifts in production itself.[81] Since in Dharm Narain's findings, the distribution of marketed surplus by size of holdings is bi-modal, this would imply that the effect of redistribution and changes in the number of landholdings on the proportion of marketed surplus would be ambiguous in general. On the other hand, Raj Krishna's

[80] Italics have been inserted by us.

[81] We deal with this question separately, when we analyze the response of production to shifts in parameters such as prices and size-class of the landholdings. We may also note here that a satisfactory investigation of the question of the effect of land ceilings and a consequent break up of larger into smaller farms requires a *general equilibrium* analysis, even within a comparative static framework, which would also take into account resulting shifts in the labor force, consumption patterns and relative prices between agricultural and nonagricultural commodities.

[82] analysis of market arrivals of rice and wheat for certain markets, by size–classes of farm output, reaches the conflicting conclusion that, except for the "very poor, and the very rich" areas, the marketable surplus (M) is linearly related to the level of garm output (Q) by the relation: $M = a + bQ$ (where the estimated 'a' involves a negative intercept). This conclusion naturally implies that shifting output between farms of different size–classes will *not* affect the volume of the marketed surplus if the number of landholdings is not changed; however, if the number is increased via land ceilings, for example, the marketed surplus would necessarily diminish (since 'a' $<$O).

Raj Krishna's startling finding, however, may be largely due to the fact that, as he had himself noted and Hanumantha Rao [65] has emphasized, he is dealing with a *single crop* and not with the overall marketing of agricultural produce by different farms. Indeed, it would appear that his data must come from farms which produce more than a single crop, or alternatively have income from different sources. As Majumdar [97] has pointed out, Raj Krishna's results imply that the output elasticity of consumption on the farm increases as output, and hence farm income, increases: a proposition which certainly is empirically untenable for India at all farm income levels.[83] Thus, it is only if these samples farms have income from *other* sources that increment in output of the foodgrain crop (rice or wheat) would not imply increment in *overall* farm income and hence would make Raj Krishna's results appear tenable.

But if then Raj Krishna's results for a *single* crop are to be considered to have come from *diversified* farms, it is not surprising that they are not consistent with Dharm Narain's results which relate to the *entire* agricultural produce from all landholdings by size–classes. If we were to assume that Raj Krishna's statistical results have validity over a wider range of foodgrains and area within the country, despite his careful caveat that they apply only to his extremely limited sample, there would clearly be important economic implications: for, from certain economic points of view, the supply of the marketed surplus of direct wage–goods such as foodgrains may be crucial but the supply of the surplus of overall agricultural produce may not be.

Apropos of this distinction between foodgrains and agricultural produce, we may also note that the precise definition of the marketed surplus is also relevant and would have to be adapted to the policy problem being considered. Thus, for example, Dharm Narain carefully states that his definition relates to the quantities that the cultivating families *directly* market. Thus, insofar as these families may themselves buy agricultural produce from the market or make payment of rent or wages in kind which, in turn, seeps into the market, the measured surplus will differ from the surplus that becomes available for *nonfarm* use. An important consequence is that, insofar as we are interested in the availability of agricultural produce for nonfarm consumption, Dharm Narain's definition will understate the marketed surplus ensuing from the larger farms which make wage payments in kind and whose purchase of agricultural produce is likely to be proportionately lower in relation to their output.

We may also remark that the Indian discussion of the marketed surplus, while it has taken different size–classes of holdings into account as a relevant variable, has not considered the possibility that al-

[83] $(M/Q)E_{MQ} = 1-(C/Q)E_{CQ}$, where E_{MQ} and E_{CQ} are the output elasticities of marketed produce and of consumption respectively and C is the consumption. Thus if E_{MQ} decreases with output, as a linear relation between M and Q ($M = a + bQ$) with a negative intercept 'a' would imply, then E_{CQ} must rise with output [97].

ternative forms of land tenure may also affect the overall level of the marketed surplus and its price elasticity. This may be via the efficiency and/or the price response of *production* varying under alternative forms of tenure (such as share cropping and peasant proprietorship). It could also result from differences in *consumption* patterns, for any given level of farm output, that may arise from differences in the distribution of the farm income among rents, wages and imputed self-incomes under alternative tenure systems. Where the tenure systems, in turn, overlap with size-classes (as, for example, when "small" farms are characterized by peasant proprietorship and the "large" farms by tenancy), the causal explanations may also overlap.

Behavior of Production

In contrast to the analysis of the marketed surplus problem, the Indian analysis of agricultural production has been empirically more systematic and analytically more interesting. As with marketed surplus, two of the major problems analyzed have been (1) the elasticity of price response and (2) the relationship of farm size to productivity. At the same time, however, the efficiency of production and investment has been discussed in the context of alternative tenure systems such as share cropping and of agrarian organizations such as cooperative farming.

(A) *Price Elasticity of Production:* There is an important, empirical distinction between the elasticity of price response of total agricultural production and of single crops or subgroups thereof. Furthermore, the elasticity of response to change in the relative price of outputs needs to be distinguished from the elasticity of response to change in the relative price of inputs to output. Each of these behavioral relationships has a bearing upon the Indian policy discussion, although it is only recently that careful distinction among these alternative concepts has begun to emerge in the policy debates.

Furthermore, practically the bulk of the systematic, empirical literature in this area has been confined to the estimation of production, or rather acreage, response of *specific crops* to changes in the relative price of *outputs*. Recently however Minhas and Srinivasan [115] have investigated the interesting question of farmers' potential response to fertilizer availability at specified prices, using crop-cutting experimental data on fertilizer productivity and *assuming* profit maximization; their work is thus aimed at predicting product response to alternative fertilizer prices but does not estimate it from observed data.[83] We should however note that the fertilizer response functions used by Minhas and Srinivasan cannot be necessarily generalized to Indian agriculture as a whole since they were obtained from experimentation done by the Indian Council of Agricultural Research and there is no reason to treat these as "typical" response functions. Hence the accuracy of predictions based on the Minhas-Srinivasan exercise is likely to be limited, despite its value in providing a systematic analytical framework for doing so.

Among the empirical examinations of the responsiveness of production to change in relative output price are: Raj Krishna's [81] estimation of *acreage* response functions for several crops in the Panjab for the pre-Partition period; Dharm Narain's [121] analysis, stopping short of econometric estimation, of shift in acreage under different crops in response to price change during 1900–39; and Venkataramanan's [173] estimation of jute areas elasticity, with respect to the relative price of jute with rice for 1911–1938. Among the other

[83] Minhas and Srinivasan also investigate the effect of share cropping on fertilizer absorption. We turn to this question later in the present section.

attempts in this direction, we should also note: Rath and Patwardhan's [144] estimation of acreage response functions for wheat for 1950–51 to 1961–62, in connection with their analysis of the impact of P.L. 480 imports on the domestic production of wheat; and Jai Krishna and Rao's [79] estimation of alternative acreage allocation functions for wheat in Uttar Pradesh for 1950–51 to 1962–63.

Other contributions of interest in this area, which depart from the focus on *acreage response to price change*, are (1) by Mann [98] who has estimated the price elasticity of *supply* (as distinct from acreage) of *cereals* (as distinct from a single crop) at 0.275 during 1952–63 in the framework of a simultaneous equations model designed to study the impact of P.L. 480 imports of wheat on domestic foodgrains production; and (2) by Hanumantha Rao [64] who has analysed the Farm Management Studies data on crop patterns in the States of Madhya Pradesh, West Bengal, Madras, Uttar Pradesh, Panjab and Bombay to find correspondence between the relative profitabilities of crops (defined in terms of income per acre) and their relative shares in the gross cropped area:[84] an empirical relationship that is compatible with, but does not necessarily follow from, the assumption that farmers respond to profit incentives.[85]

Quite apart from the differences in the period, area and crops examined, the acreage response studies by Raj Krishna, Dharm Narain and others vary in (i) their methods of estimation, (ii) the specification of the relevant price of the crop, and (iii) the selection of the relevant, *relative* price. On the method of estimation, for example, Raj Krishna as also Jai Krishna and Rao have used Nerlove–type "adjustment" models which permit separate computation of the short–run and long–run elasticities of response. The other contributors, on the other hand, have used simple, lagged regressions of acreage upon price. As for the specification of the relevant (absolute) price of the crop, different possibilities have been experimented with Jai Krishna and Rao who have tried several alternative prices—preceding year's post–harvest mean and modal prices, three month pre–sowing prices and average of three month pre–sowing and lagged three month post–harvest prices—find, for example, that their best results are with the three year averages of the pre–sowing price of wheat and consider this to be more appropriate in forming the farmers' price expectation than the *post–harvest* prices (used by Raj Krishna and by Rath and Patvardhan) or annual average prices. However, in choosing the most appropriate *relative* price deflator, the general practice has been to take weighted averages of relevant "substitute" crops. The majority of the analyses further distinguish between irrigated and unirrigated acreage, as the substitution possibilities are different for them. Further, many of them introduce overall crop acreage, rainfall and relative yields of the different rival crops as additional explanatory variables.

Among the principal conclusions to emerge from these studies is that Indian farmers vary their acreage under most crops when relative prices change. But can we really claim, with Dharm Narain, that the response is more obvious, and elastic, for pure cash crops than for cereals? The evidence on this issue does not appear to be at all clear. The elasticities of response are certainly large on some of the pure cash crops such as cotton, though not as

[84] Rao's analysis also differentiates between landholdings by size–class of holdings.

[85] Thus, consistent with profit maximization, it is easy to imagine production functions for rival crops which imply that, in equilibrium at any stated commodity price vector, the shares of these crops in total acreage are inversely (rather than positively) related to income *per acre*.

large on others such as jute and sugar cane. Further, on cereals such as rice and wheat, the elasticities estimated by these authors seem to diverge.

Raj Krishna's estimates, for example, are at 0.08 and 0.14 for irrigated acreage for *wheat*, for short-run and long-run periods respectively. In this result, he is supported by Dharm Narain's finding that, in the four major areas producing wheat in India during 1900–1939, no significant relationship could be discovered between acreage under wheat and its relative price.[86] The work of Jai Krishna and Rao, for the more recent period 1950–51 to 1962–63 for Uttar Pradesh, however, has produced long-run elasticities for wheat acreage which range up to 0.72. Similarly, for *rice*, in Panjab, Raj Krishna's estimates of the short-run and long-run elasticities are 0.31 and 0.59 respectively, indicating again a fair degree of response, whereas Dharm Narain fails to discover significant price-induced shifts in rice acreage for Bengal, Bihar, Madras and Orissa.

The divergent estimates for cereals seem to imply that empirically there is no reason to believe that the acreage response for movements into and out of cereals will be any less than that for the pure cash crops. Nor, indeed, does there seem to be any theoretical reason for such an asymmetry. Indeed, the technological constraints on shiftability of land seem to be the relevant factors in determining the magnitude of the acreage response to price change quite irrespective of the classification of the produced crops as cereals or pure cash crops. Thus, for example, Dharm Narain did find that for the Aus rice in Bengal, which directly competes with jute for acreage, acreage was indeed responsive to relative price change; and Venkataramanan has estimated the elasticity of jute acreage at 0.46.

We may further observe that if our primary interest is in the elasticity of supply or production response, the acreage response is only an incomplete guide to the total picture. The elasticity of supply response is the sum of the elasticity of acreage and productivity-per-acre responses to price change. Thus the elasticity of supply response is certain to be understated by the acreage elasticity for those crops which have no "substitute" crops landwise but where production can be increased via application of more inputs. Moreover, even if we were to consider a completely neoclassical model with factors (including land) freely adaptable to alternative crop production, a shift to a more profitable crop will generally raise its land productivity if it is intensive in the use of nonland factors and these factors are not in perfectly elastic supply to all crops taken together.

(B) *Production and Size-Class of Holdings:* The focus on production response to prices in the Indian literature has been nearly matched by the economic analysis which has resulted from the finding of the Farm Management Studies (for Uttar Pradesh, Madras, West Bengal, Bombay, Panjab, Madhya Pradesh and Andhra, during mid-1950's) that an inverse relationship obtains between farm size and productivity per acrea. This relationship, which has an obvious bearing upon the policy issues pertaining to land ceilings[87]

[86] Since Dharm Narain has not undertaken statistical estimation, and his analysis proceeds on the basis of graphical methods, comparison of his work with the econometric results of other authors can only be tenuous. Further, phrases such as "significant relationship" must be construed in a nonstatistical sense when we are referring to Dharm Narain's work. For pertinent criticism of Dharm Narain's omission of statistical techniques, and the dangers of having failed to avoid false conclusions thanks to the presence of serial correlation and an inability to face up to the identification problem, see Lipton's [91] interesting review.

[87] We may also refer to a rather different type of analytical treatment of the question of land ceilings by Khusro [76] who has ingeniously attempted to examine the efficacy of such a measure in terms of the theory of rationing.

and land grouping under cooperative farming and other forms of agrarian organization, has led to attempted explanations by Khusro [74], Mazumdar [111] and Sen [157] and further empirical work by Hanumantha Rao [66] and A.P. Rao [142], among others.

We may note at the outset that much of the important statistical evidence that is available points rather strongly towards the existence of an inverse relationship between farm size and productivity. The Farm Management Studies indeed yield this relationship, whether grouped by size-classes or taken on an individual farm basis as Hanumantha Rao [66] has done for Bombay. Independent survey for Andhra by Rao has also yielded similar results.[88]

Assuming that the results of the Farm Management Studies are statistically valid, what are the possible explanations? For, the implications for policy would generally vary with the explanation accepted. Whether higher productivity per acre also goes with greater *economic efficiency* of the smaller farms will depend critically on the explanation of the phenomenon of higher productivity.

(1) The explanation offered by Sen [157] of an inverse relationship is based on the argument that the smaller farms are characterized by peasant family cultivation and the larger farms by capitalist cultivation. Cultivation is thus carried on the small farms right up to the point where the marginal product is zero (or at least below the ruling market wage) and stops on the capitalist farms at the point where the marginal product equals the market wage. Hence the small farms have higher productivity per acre and are more efficient in the economic sense.[89]

This argument, however, raises conceptual difficulties and is also empirically untenable. If the two agrarian systems *coexist*, one may ask whether the opportunity cost of peasant family labor is not the wage that the market offers for employment by atomistic capitalist farmers. Thus, if the family is taking a decision on *overall* income derived from input of work-hours by the family as such, then will not the opportunity cost of work on *both* types of farms be equalized and the inverse relationship therefore not explained? The inverse relationship therefore will hold only insofar as we explicitly postulate that the peasant family labor cannot necessarily find alternative employment at the given wage, which is not further flexible downwards, on the capitalist farms and that the probability attached therefore to finding such an alternative employment is less than unity thus making its opportunity cost less than the wage on the capitalist farms. Mere coexistence of the two agrarian systems is not sufficient therefore for the explanation of the inverse relationship.

Furthermore, Sen's argument runs into *empirical* difficulties on two grounds. Several studies show that the small farms, not far from the bottom of the scale,

[88] On the other hand, we should note here that a few village studies, in depth, seem to throw up results which are at variance with these. Thus A. P. Rao's [142] study of data for six villages in Uttar Pradesh and Panjab shows for example that, if adjustment is made for the fallow lands and the availability of irrigation, output per acre tends to be constant by size-class of holdings. Further, C H. Shah's [161] study of the small farmers in Kodinar Taluka for 1952–53 shows that, for the principal crops: "yields per acre for the small farmers were, in the year under survey, 12.3 per cent lower compared to that of big farmers. Compared to that of medium farmers they were 7.2 per cent lower. In other words, lands in possession of small farmers were less intensively utilized to that extent." Shah does not explicitly state that this ranking would hold for *all* output and therefore it is not *certain*, although quite plausible in view of these crops covering nearly 75 per cent of the cropped area on farms in *each* category, that his results are contrary to the findings of the Farm Management Studies.

[89] The asymmetrical nature of these two agrarian systems, with respect to effects on the efficiency of production, is well known in the trade theoretic literature on domestic distortions and was earlier noted by Bhagwati [8], among others.

themselves hire labor at the margin[90] and even derive income from employment of family members in other occupations,[91] so that the opportunity cost of labor on the small farm is likely to be very real and cannot be dismissed. Moreover, Hanumantha Rao's work shows that the inverse relationship holds even when the larger (presumably capitalist) farms are ranked, so that Sen's suggested explanation is at best incomplete.

(2) Khusro [74] has noted, in particular, that the decline in productivity per acre is reduced significantly when the acreage is "standardized" on the basis of land revenue ratings (which are presumably related principally to soil fertility). Thus, one of the major explanations advanced for the inverse relationship is that the fertility of the soil is lower on the larger farms.

This argument, if accepted, raises the question whether this fertility difference is exogenous or manmade and hence, in turn, to be explained by the fact that the small farms are more efficient economically. If the fertility factor is exogenous, it could be explained by the hypothesis that the larger farms are put together by purchase of land undergoing "distress sale" and that the poorer lands are sold and the better lands retained: thus making the larger farms less fertile on the average than the smaller farms. Further, if the large farms contain an element of conspicuous consumption, the possession of land itself (regardless of quality within a range) conferring status and psychic satisfaction on these large landowners, then it could also be economically profitable for them to purchase lower quality land from the market. An implausible hypothesis is that historically the *fertile*, large farms may have broken down into smaller farms owing to a more rapid population growth. Sen, who has put forth this last hypothesis, however, forgets that this argument conceals an important indeterminacy which arises from the fact that the size of the family itself may vary by farms owing to migration or other endogenous factors and, if so, the equilibrium pattern of fertility by size-classes may not be characterized by an inverse relationship between farm size and productivity.

(3) Two other explanations related to the hypothesis of "distress sales" of land resulting in the buildup of larger farms, can be advanced. On the one hand, the enlargement of farms by acquisition of plots of land from such sales could well lead to the larger landholdings being characterized by fragmentation of the cultivated area and its being scattered over large distances, thus adversely affecting average productivity per acre and lowering that of the larger farms.[92] On the other hand, the possibility that the smaller farms are in distress could lead to their being more efficient in their use of resources, especially labor and management (the alternative being *ruination*) whereas, as Hanumantha Rao [63] has suggested, the larger farms are less efficient from the viewpoint of production, as they trade off marginal profitability against *leisure*.

(4) Another hypothesis, put forth by Khusro, is that "If there are tenurial disincentives resulting in lower input and output per acre among the tenanted holdings and if the proportion of area leased in increases with size, then the decline in output per acre with size could be partly explained by the operation of tenurial isincentives" [66]. However, as Hanumadtha Rao notes, the evidence on this issue is conflicting: while Khusro has found evidence in support of this hypothesis to

[90] This is shown by the Farm Management Studies. See Hanumantha Rao [66] and Khusro [74].

[91] This is shown by the All-India Rural Credit Surveys. For details, see Dandekar [31].

[92] This hypothesis could be readily tested through village studies designed to estimate the fragmentation of the farms by size-class of landholdings.

the extent that the proportion of land taken on lease rises as the farm size increases, the findings are just the opposite in the Farm Management Studies.

(5) A possible explanation, similar in spirit to the tenancy explanation of Khusro's, is that the larger farms are characterized by *absentee landlordship*, which results in reduced efficiency through inadequate exercise of managerial and entrepreneurial functions. While this explanation will not explain the findings of the Farm Management Studies, which relate to owner–cultivators *in residence*, it may well explain the inverse relationship in other samples.

Since many of the new hypotheses that we have suggested have not yet been tested whereas none of the traditional explanations we have surveyed appear to fit entirely any of the empirical data (wherever tests have been attempted), it is difficult not to be sceptical about the precise policy implications of this area of analysis.[93] At an a priori level, however, we may reiterate that the question is of considerable relevance to the problem of the optimal agrarian structure. In this context, we may note that Dandekar [30] has drawn upon Georgescu–Roegen's earlier work [57] to argue that, for India where there is (according to him) overpopulation in the sense that the shadow rental of labor falls below the subsistence (and hence the market) wage, the capitalist form of wage–labor organization will lead to inefficient aggregate output and the peasant family system implied by individual peasant proprietorship would be superior. Ideally, this argument would lead to an agrarian structure based on peasant families owning land in the same ratio as the overall family–land ratio. Such a view however rules out possible indivisibilities, relating to inputs, which may make *cooperation* desirable.[94] Further, as a policy prescription, it is inadequate as it does not take into consideration the economic problems of the transition from one system to the other. We may further observe that, while such an agrarian structure can be demonstrated to be statically efficient, its effect on long–run growth may be deleterious if induced savings are adversely affected and there are political limits to the governmental ability to tax (as there certainly is in India, especially with respect to the agricultural sector). This may happen via savings in agriculture being directly affected through shift in the internal distribution of income within agriculture or via income distributional changes between agriculture and other sectors as the terms of trade between them change in response to the primary improvement in agricultural output and the possible change in the consumption pattern that may be associated with the changed distribution of income within agriculture under the new agrarian structure.[95]

(C) *Tenancy, Share Cropping and Efficiency of Production:* We have already seen how the literature on farm size and productivity has led Indian economists to focus on the relative efficiency of alterna-

[93] Further, we should emphasize that the ranking by private and social profitability of the farms by size-classes may diverge from their ranking by acreage productivity. Also the static efficiency of the smaller farms, if demonstrated, may be consistent with their dynamic inefficiency from the viewpoint of savings, investment and innovation.

[94] Dandekar notes, however, that if the cooperative societies act in a capitalistic manner, they will make the system revert to the inefficiency of the capitalist, agrarian system.

[95] Again, therefore, if we are interested in the related question of what would happen to agricultural output when we have a shift in agrarian structure, we should ideally consider the problem in a *general equilibrium* framework (even if we are considering comparative static analysis). Thus, for example, a primary improvement in agricultural output due to agricultural efficiency could be overcompensated by the secondary reduction in output brought about by an agricultural price reduction induced by a consumption shift away from agriculture, thus leaving us with a net reduction of agricultural output in the new equilibrium.

tive forms of owner–cultivation: the peasant family system and the capitalist employment–for–wage system. Indian analysis of agrarian organization has also been addressed, however, to the question of the effects of tenancy, including certain important forms of it such as share cropping, on the efficiency of production.[96]

In fact, tenancy legislation has been extensively enacted in different States in India, reflecting and in turn stimulating the literature that we presently survey. Indeed, as Dandekar [29] has noted, the Indian planners have increasingly shifted their policy proposals away from the First Plan emphasis on restructuring of land holdings into efficient sized units backed by cooperative organization where scale effects make it desirable (with tenurial, land reform undertaken largely as a transitional means towards this reform of the agrarian structure), towards the Third Plan's exclusive attention to tenurial reform, inclusive of tenancy legislation.

The Indian analytical literature on tenancy has considered, among others, two principal questions of some interest: (i) where the tenant is subject to insecurity of tenure as a result of the threat of possible eviction, does the grant of permanence of tenure by legislation improve efficiency via investment in capital inputs; and (ii) are certain forms of tenancy, particularly share cropping, suboptimal from the viewpoint of efficient factor use?

Among the empirical studies aimed at examining whether the legislative grant of security of tenure to the tenants, where effectively implemented, improves the efficiency of factor use via investments which would otherwise not be undertaken by the tenant, Khusro's [72] investigation of such land reform in Hyderabad during the period 1948–49 to 1953–54 is noteworthy.[97] He hypothesizes, among other effects, that the land reform legislation would lead to a narrowing of the gap in the productivity per acre between the owner–cultivator and the tenancy groups of landholdings, presumably as the latter group improved its efficiency via increased investments induced by the tenancy reform. Khusro indeed observes that such a narrowing of the gap had occurred by 1953–54. However, the real test is whether the tenancy group had improved its productivity and it turns out that, as Dandekar [28] has noticed, the narrowing of the gap has occurred through a decline in the productivity of the owner-cultivator farmers instead. Insofar as the latter phenomenon is due to factors applicable *only* to the owner–cultivator group (as would be the case, for example, if other provisions relating to the tenancy legislation may have depressed the incentive to invest by this group of farmers),[98] then the observed narrowing of the gap would not support the hypothesis being tested. Furthermore Dandekar has pointed to the wide variations in the acreage productivity of the two groups through the period, thus mak-

[96] There is also a considerable amount of literature on the problems of (1) rural indebtedness and (2) land revenue administration, which we have not considered in this survey. Land reform discussion in India has embraced both these problems. For a reference to the major studies of the latter problem, in the light of the legislation enacted by different States such as Andhra Pradesh and Gujarat to abolish these intermediaries, Dandekar's [28] critical survey is an excellent source.

[97] Dandekar [28] has critically surveyed Khusro's study and three other similar studies, sponsored by the Research Programmes Committee of the Indian Planning Commission, to investigate the working of land reform legislation.

[98] Khusro [72, pp. 61–163] himself offers a different type of reason, also specific to the owner–cultivator group, which is of some interest: "It is well known . . . that land reforms had led to a good deal of resumption of land by owners partly because they wanted to cultivate the extra land and largely owing to expectations and psychological attitudes which this reform had led to. If the land so resumed had in fact been resumed with the intention of cultivation with at least the same standards as already existed in owner–cultivated tracts, the productivity of the owner–cultivators would have remained at least constant. . . . On the contrary it is resumed for institutional and legal reasons to safeguard

ing it unreliable to base any conclusions on a two-year comparison. On the other hand, we should note that Khusro's alternative hypothesis that the increased security of tenure would lead to a shift in the *composition* of the tenant-cultivators' investments towards investments maturing over a longer period seems to be consistent with the developments over the period.[99]

Khusro's study and other similar investigations are thus indicative but not entirely decisive in providing evidence consistent with the hypothesis relating efficiency on the tenant-farm to security of tenure and further empirical work seems called for in this area. Furthermore, we may note that the theoretical basis for this hypothesis, plausible as it seems, may be weak insofar as it is possible to argue at a purely a priori level that the implicit assumption that the cash lease tenant must finance the investments while the landlord may evict him from the lands on which he has carried out improvements may be partially or entirely invalidated by either (1) institutional arrangements for compensation to the tenant for these improvements, or (2) financing of these improvements by the landlord himself, with the return to the tenant's inputs being determined by the marginal productivity thereof. This could happen as the net payoff ensuing from such arrangements ought to induce their acceptance; and the probability of such acceptance may be high as there would be only two, readily identifiable negotiating parties involved. Whether in fact such arrangements tend to exist in practice and, if so, whether they are extensive is of course an empirical matter on which systematic evidence does not appear to have been collected over a long period and covering much of the country.

A similar theoretical objection applies to the traditional view that share cropping is an inefficient tenurial system, even when the tenancy is fully secure. We have already noted Hanumantha Rao's [67] ingenious defense of the "rationality" of share cropping when the technological possibilities of factor and product substitution are insignificant; and we have already seen how a recasting of such an explanation can reconcile the share cropping system with optimality of factor use. However, even if we were to revert to the traditional frame of analysis, we should not rule out the possibility of suitable arrangements being worked out by the negotiating parties if there is a net pay off to an otherwise blocked act of investment.

At an empirical level, in any case, no systematic attempts appear to have been undertaken so far in India to test for the alleged inefficiency of share cropping. The work of Minhas and Srinivasan [115] which we have already noted, instead tries to *predict* whether, on the assumption of profit maximization and specified technological relationships and prices of output and inputs, the share croppers (with observed shares) will have incentive to absorb fertilizers.[100] In undertaking this analysis, they are careful to note that the uncertainty of the outcome from fertilizer inputs, owing to exogenous reasons (such as weather failure or shortfalls in related inputs such as public sector irrigation) or

against future encroachments by tenants. Thus the resumer has no intention of making any investment on the land immediately or growing crops on it with the same efficiency with which he has been cultivating his other tracts. The result is to push up owner-cultivated acreage without simultaneously pushing up the production of this class." However, Dandekar has shown that this argument is not supported by Khusro's data, which register no significant change in the acreage cultivated by the owner-cultivators.

[99] This may also account partly for the fact that while the investment per acre *does* show a perceptible increase on the tenant-cultivated farms, their productivity per acre remained stagnant over the short period of the operation of the land reform.

[100] This inquiry was prompted by the shift in India's agricultural strategy, with the end of the Third Plan, towards fertilizer-intensive agricultural growth.

inaccurate application of the implied new technique, would have to be allowed for in predicting the fertilizer absorption levels at the assumed prices. This is important particularly since the application of fertilizers may lead to higher average output but greater variance of output. They also assume perfectly elastic supply of credit at a common interest rate for everybody: hence, credit is not related to farm size and status, as is probably the case in practice. Further, in making their prediction, they assume (with other analysts of this problem) that the *tenant* will be making the investments, so that the higher the crop share accruing to him the greater the fertilizer absorption. However this assumption, often made by the proponents of land reform who recommend higher crop shares for the tenant, may be empirically invalid. If the investment decisions are made, and financed, by the landlords—and politically *they* may have control over the governmental lending institutions in the rural areas, for example—then higher shares for the landlord, *ceteris paribus*, would lead to greater, rather than lower, fertilizer absorption. This issue is an empirical one and does not appear to have been treated systematically in the literature.[101]

The literature on behavioral relationships in Indian agriculture that we have surveyed so far has had direct relevance to the lively policy debate on the appropriateness of the governmental policies relating to pricing, procurement, imports and distribution of food in India. This debate has raised questions, and led to analysis, of considerable interest. It is to these questions that we now turn.

[101] Some of the land reform studies, however, have distinguished between investments made by the tenants and by the landlords. See, for example, the examination of the working of the Bombay Tenancy Act, 1948, by Dandekar and Khudanpur, surveyed in Dandekar [28].

Price Policy and Production

With respect to the effects of agricultural price policy on production, two different questions can be distinguished in the Indian debate. On the one hand, the question of agricultural prices, as such, has been discussed, largely in relation to their *stability* and impact thereof on investment. On the other hand, the question of the *relative* terms of trade between agriculture and other sectors and own–inputs has also received attention. We consider each question in turn.

(A) The desirability of having "guaranteed, minimum prices," announced prior to the sowing season, has been widely emphasised in the Indian literature.[102] Legislative and executive action in this area has, however, only recently begun: with the announcement since 1966 of minimum support prices for several major agricultural commodities such as paddy, *jawar*, wheat and maize, and the setting up in January 1965 of an Agricultural Prices Commission to assist in formulating these prices and the Food Corporation of India in making the necessary purchases to make these prices effective where necessary.[103]

The underlying theoretical basis for the guaranteed minimum price approach seems to have had numerous elements. (1) Dantwala [34], among others, has referred to the "insurance" aspect of such a policy and its resulting elimination, via the provision of a floor price, of that aspect of uncertainty which might deter investment. Whether, however, open market operations in the agricultural market, designed

[102] Cf. Dantwala [34], the Report of the Agricultural Prices Commission on Price Policy for Kharif Cereals for 1965–66 Season [176] and the Report of the 1966, [Venkatappiah] Foodgrains Policy Committee [177].

[103] As happens with many such bodies, the Food Corporation of India has managed to multiply its activities well beyond this area and has even involved itself in fertilizer distribution. Cf. the Foodgrains Policy Committee Report [177].

to mop up supplies when the price tends to sag below the floor price, is a preferable alternative to an insurance scheme which, among other differences, does not involve direct State trade in agriculture has not been debated in the Indian literature. (2) The notion that the minimum guaranteed price, on the other hand, is part of a *stabilization* policy aimed at evening out fluctuations has also been explicitly developed by other economists, including Dandekar [32]. However, the distinction between price and income stabilization for agriculture or foodgrains has been made all too rarely.[104] Nor has the problem raised for buffer stock operations (on which the proposed stabilization measures must rely) by the *trend* rise in agricultural prices been discussed.[105] In turn, the critical question as to whether private speculation itself tends to be stabilizing or destabilizing in the field of agriculture in general, and specific foodgrains in particular, has not received the attention it deserves. The view that private trade is destabilizing (in some sense) and inefficient (in eliminating spatial price differentials, among other things) seems to have been widely accepted as obvious. In this connection, at least two empirical investigations are of interest. Venkataramanan [174], who has examined the data on spot and futures price quotations at the East India Cotton Exchange, which of course represents a fairly developed market, and stocks of cotton in Bombay, has found that the Keynes–Hicks theory of "normal backwardation" is consistent with the observed facts. Moreover Uma Lele [90], who has studied the sorghum trade in Maharashtra State, for five primary markets in Sholapur district and two terminal markets, has found that much of the regional price differentials (where not illusory and accountable by differences in grain quality) can be accounted for by factors such as transportation bottlenecks, freight costs and governmental restrictions and bans on movements. (3) Finally, nearly all economists writing in this area [32] [34] have expressed the view that, in addition to the "insurance" element, the guaranteed minimum prices should include a margin intended to "help in assuring the progressive farmer that additional effort and expenditure for the purpose of increasing output will bring him an adequate return." [177, 54]. This view amounts to arguing, *ceteris paribus* for improved terms of trade between agriculture and other sectors,[106] which would permit the (relative) influx of resources into agriculture. What would be the optimal policy for bringing about a shift in the agricultural terms of trade and what would be their optimal level are issues which this line of policy analysis opens up. These are also the issues which belong to the second class of questions, relating to *relative* agricultural prices, to which we now turn.

(B) Much of the Indian literature and debate has inevitably dealt with this class of problems, although the discussion has

[104] Dandekar [32, 27] notes the difference but opts for a rather strange solution which aims at *both* price and income stabilization for foodgrains, without considering any alternatives: ". . . , in the interest of the producers, any measures of stabilization of prices such as through operating support and ceiling prices, should be accompanied by measures of income stabilization through appropriate credit and insurance policies."

[105] In deciding on the optimal level of the buffer stock, it would be necessary also to consider the possibility of holding free foreign exchange reserves, with pipelines set up for activating imports when necessary, since instability is likely to arise in respect of commodities other than foodgrains as well. A decision to hold buffer stocks for *each* item (including agricultural produce and foodgrains) where instability will arise may be suboptimal and the holding of foreign exchange reserves instead, or some combination of the two measures, may be superior. This question has been neither posed nor explored in the Indian literature.

[106] This should include improvement in the relative price of agricultural output vis-a-vis agricultural inputs.

been confined to the question of whether Indian agriculture has been subjected, over the first three Plans, to a *trend* situation of price disincentives. Much of this debate has centered on the behavior of the agricultural terms of trade in general, although the foodgrains terms of trade have been distinguished. Furthermore the question of input prices and the net burden of taxation on the agricultural sector vis-a-vis other sectors has also been raised.

Dantwala [34], in an important contribution which surveys the entire range of governmental policies over the period of the three Plans, has critically examined the prevailing view that agricultural prices have had a strong disincentive element. He finds this view inconsistent with the facts insofar as the recorded time series of the agricultural terms of trade fail to register a deterioration over the period.[107] In fact, relative stability over the period, with the exception of a sharp dip in 1955–56 (which led to a rather slow stepping up of public sector investments), seems to have characterized the terms of trade between agricultural and nonagricultural commodities. On the other hand, if we examine the terms of trade between *foodgrains* and nonagricultural commodities, or between foodgrains and the overall index of wholesale prices, there is certainly evidence of a more distinct deterioration during the first Plan period which is eliminated towards the end of the period.

While these facts are interesting in themselves, they beg the more relevant question as to *which* level of the terms of trade, whether between agriculture and the rest of the economy or between foodgrains and other commodities, should be considered *optimal* and whether, in relation thereto, the recorded terms of trade for agriculture or foodgrains were "unfavorable." Dantwala raises this issue tangentially when he argues that, in relation to 1939 prices, the cereals index was already 444 (with 1939=100) and the general price index only 380.6 in 1952–53, the baseyear of the new price indices. Hence, *in relation to the 1939 terms of trade*, the evidence over 1951–1966 indicates "favourable terms of trade for agriculture."[108] Similarly, Dandekar's [32] contention that the terms of trade for cereals show distinct improvement during the Third Plan largely in the drought years at the end is also implicitly raising the same unanswered question as to the optimal level of the terms of trade.

It is interesting however that the participants in this debate have not come to direct grips with the fundamental question of determining the *optimal* level of agricultural prices vis-a-vis other prices. Thus, for example, there has been no attempt at determining how these internal terms of trade compare with the international rates of exchange between agricultural and other commodities and whether exchange rate and trade policies conferred an excessive, in the sense of suboptimal, incentive for resources to flow into the nonagricultural sector.

At the empirical level, however, greater sophistication has been introduced into the discussion by examination of the relationship between agricultural output and input prices and the net burden of taxation on agriculture relative to other sectors. Dantwala [34], after making the valid point that little empirical evidence is

[107] Dantwala [34] makes the important observation that: "... For some commodities like cotton, there has been a statutory ceiling on prices, and though in reality the ceiling has never been operative, the office of the Economic Adviser which prepares the index series records only the ceiling prices. Thus, for commodities like these, the index number under-estimates the rise in prices."

[108] Dantwala [34] then notes that "This would perhaps, explain why in January 1957, when the [new] cereal price index stood at 95, the Government of India thought it fit to set up a high powered committee "to examine the causes of the rise in prices and to suggest remedial measures."

available and, where available, it does not indicate a high elasticity of response of agriculture *as a sector* to its terms of trade with respect to other sectors, proceeds to examine the available information for input/output prices for Assam, Panjab, Kerala, Orissa and West Bengal and finds conflicting evidence in relation to 1939, while noting the unusually unreliable character of these series. The information on the relative tax burdens, however, is sound and Dantwala quotes Ved Gandhi's [56] thorough work to show that sectorwise agriculture has received exceptionally favorable treatment. This is particularly because of the direct agricultural taxes amounting on the average to no more than 2 per cent of the value of agricultural production during the planning period.[109]

Price Policy, Distribution and Imports

Regardless, however, of the issue as to whether Indian agricultural production was discouraged by governmental failure to provide the optimal terms of trade, the question persists as to whether the entire set of governmental policies, designed to deal with a continuing situation where at constant prices *foodgrains* production was short of the demand fed by income and population growth, were optimal. This question, in turn, has provoked a considerable amount of controversy.

The governmental policy package has essentially involved reliance on largescale P.L. 480 imports to supplement overall supplies of wheat and distribution thereof through a public sector system of fair price shops. Further for wheat, and more so for rice where there has been no equivalent P.L. 480 program, internal procurement of foodgrains from the producers has been attempted. Furthermore, since 1964, the country has been divided into several food zones which rule out interzonal free private trade: and the perpetuation of this system has been largely defended by reference to the government's procurement and distributional policies.

The policies just described have been severely criticized by Indian economists. However, while there is general agreement that the governmental procurement of internal foodgrains for public sector distribution to the low income groups was totally inadequate and the reliance instead on P.L. 480 imports for this purpose was excessive, the critics have been divided on almost everything else. Thus, for example, Raj [137] has argued that the zonal system has accentuated the reliance on imports whereas Dantwala [34] and the Foodgrains Policy Committee (which included D.R. Gadgil) [177] have contended that the zonal system facilitates greater procurement, implying that, *ceteris paribus*, it reduces reliance on imports. The zones have also attracted considerable controversy in relation to other issues such as their impact on economic efficiency and political integration: Raj Krishna [84] [176] and Raj have been among the principal critics.

(1) The view that the reliance on food imports was excessive and that India could and should have managed without P.L. 480 imports has been variously argued. Raj [137] has argued that imports could have been moderated, even eliminated, as there was enough foodgrain to go around "if distributed equitably."[110] While this is

[109] Quite aside from the fact that politically it is difficult to tax the agricultural sector, when the bulk of the votes are in that sector, there may be another problem here. From an income distributional point of view, the agricultural sector possibly has a relatively larger proportion of its income originating on the small farms belonging to an income level which cannot be taxed on equity criteria. On the other hand, this implies that taxation (which is necessarily not lump-sum), and hence incentives on that account, will be biased against the non agricultural sector, *ceteris paribus*.

[110] Raj refers to nutritional standards to arrive at an average per capita consumption figure of 13½ ounces as minimum cereal intake and finds that "except in two years (1951–52 and 1952–53), it would have been possible to ensure this from domestic production alone."

a correct statement of fact, it does not rule out the existence of excess demand at a given price for foodgrains, and hence the important question as to whether imports *should* not after all have been permitted (or sought, under the aid program) to moderate a rise in the price. In assessing this question, it is necessary to remember at least two pertinent points: (1) if excess demand for cereals were to be diverted, thanks to rationed distribution for instance, this demand could spill over into other consumption (instead of turning into involuntary savings) and, in turn, cut into exports, for example, and thereby affect the foreign exchange position much as imports of foodgrains would; and (ii) if acceptance of P.L. 480 imports led, in the ultimate analysis, to a *greater* total inflow of foreign assistance, this in turn would be a positive factor in favor of such a policy, *ceteris paribus*. In short, whether self-sufficiency in foodgrains is an acceptable objective of short-term or long-term agricultural policy is itself an issue which must be assessed in the light of a general equilibrium analysis of the entire economic position, including aid flow sensitivity to alternative policies, instead of being regarded as axiomatic.[111]

Raj Krishna's [84] indictment of the governmental failure to step up internal procurement, while imports under the P.L. 480 program continued, raises similar questions. It is indeed true that the facts on imports and local procurement of grains show a greater amount of procurement in the first Plan than in each of the subsequent Plans (when P.L. 480 imports became available), despite the easier food situation during most of the first Plan. And it is also correct to maintain that such a policy violated the public pronouncements with respect to the achievement of self reliance in foodgrain availability. On the other hand, it does not follow that the policies actually followed were suboptimal if one assesses them in terms of economic efficiency rather than in relation to self-sufficiency as an objective.

(2) We have already noted that the question of the effect of zonal arrangements on the food deficit and import levels has been raised, in this connection, by Raj. In fact, this issue leads us directly into the entire range of questions relating to the economic efficiency of zonal arrangements and their role in a national foodgrains policy.

The zonal arrangements in India sprang up largely thanks to the action of the so-called "surplus" States such as Andhra Pradesh, Panjab and Madras whose primary motivation appears to have been to maintain artificially low prices (in a situation of rising prices) within their boundaries by curtailing the normal outflow of grain through private, interState trade. This phenomenon raises the natural, but unexplored question as to whether the rural interests in the "surplus" states, which are thus being denied the advantages of more favorable terms of trade, are really less influential politically than the urban consumer groups to whose interest the zonal policies appear to cater. Two possible explanations, however, may be worth exploring. (i) On the one hand, it is possible for the more influential, larger landlords to make greater profits under

[111] Dantwala [34] makes a similar point when he notes, with respect to the possible adverse effect of P. L. 480 wheat imports on wheat production, that: "The major component of P.L. 480 imports was wheat and it is reasonable to assume that these imports affected the prices of wheat or at best also of other substitutable cereals from the consumer point of view, but could not have had much impact on the prices of commercial crops. The expected consequence of this relative shift in prices in favour of commercial crops would be a shift in agricultural inputs for their production. Assuming that this is exactly what happened, would such a development be necessarily injurious to Indian agriculture or the Indian economy as a whole? It is, of course, true that higher foodgrains production is very vital to India's economy, but a stimulated growth of non-foodgrain crops is of no less importance for the overall national economy . . . ."

zonal arrangements by getting access to scarce, State–distributed licenses to export their output of foodgrains to deficit States which have higher prices under these zonal arrangements than otherwise. Such a practice also redounds to the benefit of the politicians who thus develop another area of patronage and possibly even direct profit to themselves. There is some evidence that this explanation might have relevance in Andhra Pradesh. (ii) An alternative explanation may be that the political situation is based in the States on a balance of urban and rural interests. Thus, while food prices are kept low by zonal restriction on the ouflow of grains, the not–so–poor farm groups are "compensated" by the provision of negligible tax rates on agricultural income.[112] The danger in this kind of politico–economic pattern is that ultimately the whole operation would imply that the marginal tax effort, for investment and other purposes, would have to come to rely on the extremely narrow base provided by non-agricultural, urban classes outside of the group sheltered by the Fair Price Shops.

While, however, the zonal system has originated in the actions of the "surplus" States, it has found some distinguished supporters among the economists despite the severe criticism direct at it by several economists. The defence of the system is best summarized in the Foodgrains Policy Committee Report [177]. Arguing that the interState movement in foodgrains should be undertaken only through State operations, and that zonal restrictions on private movement should continue, the Committee have claimed the following advantages for such an arrangement:

> First, this is necessary for ensuring equitable distribution to different States; trade, if untrammelled, would tend to move the surpluses of one State to points of highest purchasing power in another and not to those of greatest need. Second, it would enable Government to keep prices at levels, which are reasonable for both consumer and producer; private trade, by catering for the well–to–do consumer, would be in a position to push up prices, if allowed to compete with Government. Third, if the trade is allowed to purchase within the State and sell outside it on its own account, it would not be possible to ensure maximum procurement by Government and Government agencies.

These arguments, however, are untenable. While it is true that the market system will not in itself correct an undesirable income or consumption distribution, it is a nonsequitur to deduce that the optimal way of achieving a desired distribution is to eliminate the market system and substitute governmental trade instead. The second argument is also a distributional one insofar as we can make any sense of it, and subject to the same criticism.

The final argument, which constitutes really the central defense of the zonal arrangements, is incomplete, even if factually correct,[113] and must be dismissed if the zonal arrangements are looked upon from the viewpoint of economic efficiency. In order to appreciate this, it is necessary to examine the main features of the policy package advocated by the Foodgrains Policy Committee.

They have recommended that procure-

[112] Similarly, the deficit States "compensate" their urban groups by distribution of Central supplies of imported foodgrains through Fair Price Shops at subsidized prices.

[113] The argument of the Committee that, thanks to the zonal arrangements, the governmental procurement of rice during the 1965–66 season has been higher than during the 1964–65 season, "in spite of a reduction of 17 million tonnes in foodgrain production caused by drought," surely cannot be accepted at its face value. Indeed, the very fact that there was a drought and hence a great demand for fair priced cereals during 1965-66 might have prompted more procurement of rice (as distinct from wheat, where we may note that P.L. 480 imports were undertaken instead); whether the procurement would (as also *should*) have been more or less if the zonal arrangements had not existed remains an open issue.

ment of foodgrains must be undertaken, apart from buffer stock purchases, for distribution at fair prices to certain classes of low income consumers in both rural and urban areas. Furthermore, they have argued that the procurement must be undertaken at prices *below* the market prices in order to prevent the government having to find the resources for financing the subsidy that would otherwise be entailed. Thus the Committee's recommendations effectively involve subsidizing the foodgrains consumption of certain low income groups and financing this subsidy by taxation of the farmers producing these cereals, this taxation being implicit in the fact that procurement would be at less–than–market prices. In this context, the restrictions on private interzonal trade are looked upon primarily as a means of making this procurement "easier," presumably because the apparent element of taxation would be smaller, *given* the fair-price at which procured foodgrains are to be sold, if the open market price in the "surplus" States (where procurement will presumably be carried out) is kept lower by ruling out interzonal private trade.

This view of the policy package, which seems to rationalize the zonal restrictions, is however open to serious objections. There are two particular aspects of this policy which are controversial: (i) the method of subsidising the low income groups; and (ii) the method of financing the subsidy.

Concerning the former question, it is not clear that an outright financial subsidy to the specified low income groups, index-linked to the cereals price index, may not be a less expensive system than a distributive system based on Fair Price Shops, governmental trade and distribution.[114] In assessing this question, we would have to consider (i) the relative efficiency of a public distribution system, from the viewpoint of waste in storage for example;[115] (ii) the administrative costs and feasibility of either system; and (iii) the possible, though not necessarily considerable, advantage of having a State distributional system which can be readily exploited to handle sudden emergencies such as the Bihar Famine of 1967, when an enormous operation to shift foodgrains to this area had to be mounted.

The question of the optimal method of financing the subsidy to the low income groups raises still more complex issues which have not received the attention they deserve. Before we discuss these alternatives, however, we may note the objections to the Foodgrains Policy Committee's assumption that, if one is to levy an implicit tax on the farmers to pay for the subsidy, zonal restrictions make this task "easier." Raj Krishna has argued, in a brilliant note of dissent to the Agricultural Prices Commission's 1965–66 Kharif Cereals Report's similar ideas, that this view focusses merely on the fact that procurement in the surplus States (by the Centre) would be cheaper but ignores the fact that *more* would have to be procured since, with interzonal private trade removed, "the responsibility of meeting the *entire* deficit of deficit States falls on the Central Government." [176, 38]. Raj Krishna has in mind the possibility that politically the deficit States would have to be "compensated" for the eliminated, private inflow of foodgrains. But even if we rule out such a direct "compensation," the problem remains. For as open market prices in the deficit States rise to levels higher than what they would have been if zonal restrictions were eliminated, demand for

[114] Identification of recipients eligible for the "dole" would be as difficult or easy as their identification for a ration card.

[115] Sundaram's [168] careful analysis of the suboptimality of the existing P.L. 480 landings at different Indian ports, given the ultimate destination points, is also of relevance in assessing this issue.

foodgrains would be diverted to the Fair Price Shops, thus raising the offtake from these shops and hence also the need for procurement. This would happen insofar as those entitled to access to these shops are not already utilizing it fully in the no-zonal-restrictions situation: as is indeed likely to be the case.[116] Furthermore, even if this were not the case, the rise in the open market price level could certainly lead to politically effective demands to let more income groups have access to the Fair Price Shop system facilities. However, against this, we must balance the fact that, in the surplus States, demand would be diverted *away* from the Fair Price Shop system; hence the *overall* need for procurement, to service the Fair Price Shop system, may reduce rather than increase—a possibility which is ignored by assuming unrealistically that the Fair Price Shops system applies only to the deficit States. The question as to whether procurement will be "easier" under the zonal restrictions system is thus a complex one and cannot be answered unless the analysis takes into account the strength of the low income groups and demand diversion to, and away from, Fair Price Shops as a result of the zonal restrictions in *both* surplus and deficit States: a priori, it is impossible to rule out the possibility that the quantities to be procured will increase under the zonal system.

But, even leaving this question aside, the zonal arrangements conceived as an instrument for securing procurement involving a hidden tax element represent a method of levying taxation that is ethically inequitable, economically inefficient and politically injurious to national integration. Raj Krishna [176] has correctly pointed out that there are "surplus" farmers and not "surplus" States. A policy which aims at concentrating tax-element-inclusive procurement in surplus States, while ignoring the fact that prosperous or surplus farmers exist even in deficit States, is an inequitable one. Moreover, the tax-element-inclusive procurement of foodgrains which are demanded by the low income groups is economically inefficient insofar as it discriminatorily taxes farmers who happen to be producers of these specific commodities and thereby pulls away resources, *ceteris paribus*, from the production of these commodities. Even if tax-element-inclusive procurement is considered to be the only *feasible* method of taxation,[117] to finance the subsidies for the low-income groups, there is no reason why it should be confined to the commodities which happen to be demanded by the low income groups.[118]

Finally, the zonal system, on which such a procurement system is grafted, must inevitably lead to political disintegration. The cynical reluctance of the surplus States to let their grain be procured for transfer to Bihar during the 1967 famine is only an extreme example of the inward looking approach of these States to a national food policy. A condoning of the

[116] There seems to be evidence that only the extremely low income groups generally utilize the Fair Price Shops even though more groups have the right to do so.

[117] This is often asserted in the Indian policy debates. However, it is by no means clear that where procurement has succeeded, the price paid has always included a tax element or, when such tax element is present, an alternative form of agricultural taxation would not have been feasible (and possibly preferable on efficiency grounds). This is an important, neglected area of empirical research. We may also mention in this context, while we are essentially discussing alternative forms of agricultural taxation, that Dharm Narain [119], Sen [154] and Bhagwati [4] have discussed the Preobrazhensky-type problem of how the terms of trade between agriculture and industry could be adjusted in order to extract a "real" surplus from agriculture to "finance" capital formation in industry.

[118] We may also note the further point, made by Khusro [75], that, even if the zones were not accompanied by tax-element-inclusive procurement, they would have adverse allocational effects by discouraging the production of foodgrains in the surplus States (which presumably have comparative advantage in such production) and encouraging it in the deficit States.

zonal system would only accentuate these fissiparous tendencies. In fact, as Raj Krishna had predicted, the zoning system has spread to *within* States, with districts turning into *de facto* zones in States such as Madras and Kerala. The argument of the Food Policy Committee [177] in this connection is interesting:

> Another criticism of the restrictions on inter-State movement of foodgrains is that they undermine the unity of the Nation. We do not consider this to be a valid criticism. The system does not envisage a ban on the movement of the surplus from the surplus States to the deficit States. What the system implies is that the inter-State transfers will be effected on a regulated basis by a public agency which is amenable to social control and discipline. In a situation of overall shortage, if inter-State movement of foodgrains is allowed to be undertaken in an unregulated and uncontrolled manner, it would indeed create scarcity conditions in the relatively poorer regions of the country. Such a development can have a far more damanging influence on the unity of the Nation.

This argument, however, has little practical relevance. In practice, for the very reasons that the surplus States have pushed for zonal restrictions, they have frustrated the Central Government's attempts, via the Food Corporation of India, to procure foodgrains for shipment to deficit States and have generally forced the Central Government to resort instead to P.L. 480 imports for such supplies (thus lending substance to Raj's argument that the zones have led to increased imports of foodgrains). To argue therefore for a zonal policy, knowing fully well that the chief supposed advantage from it contradicts the very purpose for which it is politically designed and adopted, is somewhat naive and has inevitably, even if unwittingly, strengthened the interests opposed to a truly national food policy.[119]

In fact, it is significant that the move to abolish, or at least enlarge, food zones to include *both* deficit and surplus States in single zones, has come from many surplus States themselves during the bumper crop of 1967–68. With prices sagging in these States, there has been a reversal of their attitudes: the producer pressure groups appear to have become more important and have sought freer access to the deficit State markets. At the same time, procurement has been permitted only at exceptionally favorable prices, leading to an exasperated critique of politicians by the leading zone supporting economist, Dantwala [36]. Clearly, the political assumptions that the zonal system would permit procurement at *tax-element-inclusive* prices, in the interest of a *national* food policy, have been shown up to be, at best, tenuous. However, instead of taking this opportunity to eliminate the zonal restrictions altogether, the Central Government is now in the role of zone supporter, opposing several surplus States' desire to let the zones widen or perish.

Existence of Surplus Labor or Disguised Unemployment

Prior to concluding our survey of the Indian literature on agricultural policy, we must examine the important issue as to whether there is surplus labor (or disguised unemployment) in the Indian economy. The assumption that this is indeed the case has formed the basis, as we have already seen, for much analytical thinking in India.

[119] In this connection, however, we may note (1) that the Central Government's ability to force the surplus States to fall into line and eliminate zonal restrictions might have been seriously inhibited by the fluid political situation within the Congress Party and the critical role played in Prime Ministerial successions, by the Chief Ministers of these States; and (2) the fact that, even if these zonal restrictions were abolished, we could not have ruled out altogether the imposition of numerous clandestine restrictions, on export of foodgrains, by the recalcitrant, surplus States.

Although the existence of surplus labor is regarded as almost self-evident by many Indian economists, and was discussed in early writings of economists such as Bhabatosha Datta, the unorthodox view of Schultz [151] who has used Indian data to argue that the phenomenon does not exist, has prompted renewed interest in the subject. In reviewing this literature, we must begin by differentiating among the numerous alternative definitions of, and hence presumed evidence in support of, the presence of surplus labor which are to be found in the literature in this area.

There are many alternative definitions, sometimes explicit but often implicit, of "disguised unemployment" in the literature, which do not necessarily coincide in scope even within the context defined by the Indian economic and institutional structure. (1) We have the definition due to Arthur Lewis which defines disguised unemployment as a situation under which it is possible to get a supply of labor from agriculture to the industrial sector at a constant real wage. (2) We also have the definition of disguised unemployment as a situation under which the social marginal productivity of labor in a sector such as agriculture is less than the wage rate at which labor can be hired: the wage rate is inflexible downwards because of the biological subsistence requirements. (3) Disguised unemployment has also been defined as a situation where the private marginal productivity of labor is zero in agriculture, so that the withdrawal of labor from agriculture would result in a fall in agricultural output. Even here, it is necessary to make a distinction between a *ceteris paribus* withdrawal and a *mutatis mutandis* withdrawal, as these two alternative varieties of withdrawal would lead to different effects on agricultural output in general. (4) Finally, disguised unemployment may be defined simply as a situation where, given the social objective of maximizing the value of current income, the combination of techniques and resources is such that the shadow wage, and hence the social marginal productivity (SMP), of labor is zero. Our analysis will be concerned with this specific definition: we will review the Indian literature on surplus on the assumption that the objective of the analysts is to discuss and test the proposition that, in the Indian context, the SMP of labor is zero.

Note first that zero social marginal product (SMP) will not necessarily involve zero private marginal product (PMP). If we assume a single sector (e.g. agriculture), a peasant family system of farming and a system of allocation of labor time which involves maximization of family-group income (even if the division of product may be on different principles), zero social marginal product will naturally lead to zero private marginal product. On the other hand, if we were to assume a capitalist system of farming, where landless labor is hired for a wage which institutionally exceeds the zero shadow wage, we would observe a positive marginal product (which would equal the market wage).

Similarly, zero SMP will not *necessarily* imply that, if the labour force were reduced in a sector, that sector's output would fall. Thus in a model where the real wage of agricultural labor is institutionally fixed in terms of a constant utility level derived from consuming both agricultural and manufactured goods, factors are immobile between sectors, there is capitalist farming in agriculture, and the SMP of agricultural labor is zero (with income distribution keeping the unemployed alive), a reduction (say, by influenza) in the agricultural labor force would have no primary impact on agricultural output. On the other hand, it would imply that the expenditure otherwise made by the deceased labor force would now be made by others. If, as a result of this implied income redis-

tribution, the demand for, and hence the relative price of, manufactures falls, we would then have a reduction in the binding nature of the institutional wage constraint and hence greater output of agriculture in the new equilibrium. Conversely, it is possible to show that, even when there is positive SMP in agriculture, the effect of reduction in the labor force may well be to maintain agricultural output constant.

Furthermore, zero SMP does not imply that the supply of labor from the sector where this is so will necessarily be perfectly elastic at some real wage. Thus, for example, if zero SMP (and zero PMP) obtains in a peasant family agriculture, with individual rather than group income maximization such that each individual will equate his average product on the farm with his marginal product in manufactures (a la Arthur Lewis), then successive supplies of labour to manufactures will raise the average product on the farm, consistent with zero PMP and SMP in agriculture continuing, and thus the marginal cost of labor supply from agriculture will continually rise (instead of being constant).

Finally, we may note that the common assumption that surplus labor must be in the agricultural sector, leading to predictions such as the elastic supply of labor to the nonagricultural sector and the constancy of agricultural output as labor moves out of agriculture, is itself restrictive. In essence, we can think of labor carrying a zero shadow wage for the economy *in toto*. If we look at the empirical situations, it is not unrealistic to postulate an economy with a common, institutionally determined wage (which exceeds the zero shadow wage) at which employment is undertaken in capitalist agriculture and capitalist manufactures. In practice, it is also possible to find in fact the coexistence of "family" and "capitalist" modes of production in *both* agricultural and urban areas: so that, in this instance as well, zero SMP may obtain with respect to *all* sectors.

The institutional features of an economy thus have a critical relevance to the manner in which zero SMP "accommodates" itself in the system. Hence, the "tests" and "measures" of surplus labor, which have been devised in the Indian and other contexts, have to be treated with great care.

(1) Thus, for example, it has been argued that there cannot be surplus labor in India because labor is hired at a positive wage in all farms, whether small or large. This argument presumes implicitly that the surplus labor is to be found on the peasant family farms and ignores the possibility that capitalist hiring of landless labor at an institutionally determined wage on all farms is compatible with zero SMP.

(2) Schultz's [151] famous test, on the other hand, has proceeded along a different route. He takes the influenza epidemic in India during 1818–19, arguing that the sudden and significant reduction in the labor force that it entailed provides a laboratory type experiment to discover surplus labor in India. On finding that agricultural acreage (and output) declined in consequence, in the year 1919–20, Schultz concludes that labor was not in surplus in agriculture.

Schultz, however, has another supplementary argument at this stage. He hypothesizes an agricultural production function of the following type:

$$Q = A \cdot (L)^{\alpha} \tag{34}$$

where

Q = output
L = labor force
A = technological constant
α = "labor coefficient"

Arguing that certain unpublished sample studies indicate the value of the "labor

coefficient" to be 0.4, Schultz further fits a regression equation on the data for the reduction in acreage (taken as proxy for output) and in labor force during 1919–20 in ten different States to find that the indicated labor coefficient is 0.349, and the hypothesis of 0.4 lies well within the confidence interval based on twice the estimated standard error of this estimate. Schultz further seems to derive greater confidence in this coincidence (between the values of the labor coefficients in his regression and in the unpublished sample studies) because a study of the share of agricultural income going to labor in Panjab during 1947–48 yields the figure 0.34 which happens to be consistent with the competitive implications of the hypothesised production function for agriculture. How does this argument strengthen the first argument which depends exclusively on showing that agricultural acreage (output) declines with the decline in the labor force? Clearly, if Schultz can produce evidence that the data for the ten States are consistent with the hypothesized production function with coefficient $\alpha>0$, then he can argue that surplus labor *cannot* exist at all since labor would *always* have a positive (social) marginal product in agriculture. Thus, the second argument is aimed at a stronger hypothesis (namely, that surplus labor cannot exist at all in India, with the given technology) than the first argument (which would only show that, for the range of variation in the labor force which the influenza epidemic entailed, there was a decline in output and hence there was presumably no surplus labor).

At the empirical level, Schultz's argument is tenuous on at least two grounds. (i) The coincidence of results at different points of time (such as 1919–20 and 1947–48) and for different parts of India (which are not exactly integrated in terms of their land or labor markets) does not necessarily reveal a regularity. (ii) Furthermore, Sen's [159] recalculation of the Schultz regression, adjusting for three omitted States and for errors in estimation of the labor force, yields a labor coefficient which does *not* coincide with the coefficient in the unpublished sample studies and the estimate of labor share in agricultural income. On the other hand, if Schultz is willing to concede the irrelevance, to his 1919–20 test, of the evidence produced for sample villages and different periods, it would be possible for him to contend that his evidence is consistent, as it stands, with the hypothesis of an agricultural production function of the type: $Q=A\cdot(L)^{\alpha}$ $(\alpha>0)$, so that the Indian experience during the influenza epidemic (even when corrected for statistical errors and omissions) is consistent with there being no surplus labour (for any labor-level altogether). In this connection, it is relevant to note Harwitz's result, cited in Schultz [152], that if these Indian data are examined for the null hypothesis of zero marginal product of labor, the null hypothesis is rejected because "the observed data have one or two chances in a hundred of having come from an uncorrelated population, under the rather conservative test of the null hypothesis"; and Schultz's conclusion is valid (only) up to a 5% level of significance on this test.

Furthermore, there are serious objections to Schultz's use of the influenza epidemic as an *experiment crucis*. (1) The influenza epidemic naturally raises doubts (considered to be unimportant by Schultz [152] in light of "medical judgments" obtained, but contested by Reports written at the time) about the debilitating effects on those who survived. This doubt is particularly enforced when we recognize that the epidemic continued, in some degree, into 1919–20 itself. (2) It is also a matter of judgment, left unsettled by Schultz, whether the lapse of just a year, with some continuation of epidemic conditions, was adequate to obtain an adequate test of

whether the "disorderly conditions" still continued and therefore a later year might not have provided a better guide to the required comparison.

Quite aside from these two rather obvious objections (both anticipated by Schultz), which render the conclusions drawn from the experiment fairly tenuous, there are two major critiques which can be advanced against Schultz's conclusions. (3) Shakuntala Mehra [112] has made the significant statistical finding that, if we break down the post-epidemic agricultural year into the two major Indian harvesting seasons, *rabi* and *kharif*, then there is ample evidence that the immediate harvest (*kharif*) after the epidemic registered no significant decline in the output (acreage) level, whereas the decline was concentrated in the later, *rabi* harvest (which, in any case, fluctuates widely owing to seasonal factors). It would appear therefore that Schultz's inference that Indian agricultural output declined with the influenza epidemic is, at best, dubious. In any event, Schultz's failure to take the two seasons into account, and the absence of systematic quantitative analysis of the two harvests, make it impossible to attach any significance to Schultz's conclusion that the Indian agricultural output declined with the epidemic. (4) There is also an analytical difficulty with the argument that decline in the agricultural output with the decline in the labor force implies that surplus labor is zero: the argument is just a nonsequitur, as a general proposition (as we have already seen). Thus consider the case where there is peasant family farming, co-existing with capitalist farming. Let the total labor supply be such that the shadow wage of labor is zero: so that we have "surplus" labor in its fundamental sense. Let further the peasant family work under the rule that the average product of an individual member is equated with the marginal product on the capitalist farms: and let the PMP on the peasant family farms be zero. If then the labor force declines on the peasant family farms, due to influenza, but the shadow wage of labor still continues to be zero, we would have a higher average product on the family farms, therefore a higher real wage at which labor will be employed on the capitalist farms, therefore a reduction in the agricultural output on the capitalist farms, and hence a reduction in the total agricultural output. Thus, we have again shown the compatibility of a zero shadow wage for labor and decline in the agricultural output as the population (labor force) declines.[120] Nothing can be concluded, therefore, about the existence of "surplus" labor without a careful investigation of the institutional structure of the sector within which surplus labor is assumed to inhere.

(3) There have also been *direct* measures of surplus labor in India, following the classic methods of Paul Rosenstein-Rodan

[120] Admittedly, this result depends on the twin assumptions (1) of the division of the given supplies of land (and other factors) into two groups: peasant family farms and capitalist farms, and (2) that the individual, rather than the group, on the family farm maximizes income. If we were to relax the second assumption, for example, and assume that the group maximizes income, then the market would equalize the marginal products on both sets of farms and, with the supplementary assumption that each peasant family has sufficient income for subsistence from own-farming, the wage could fall to zero with zero SMP. In such a case, the Schultz test would be perfectly adequate, of course: zero SMP would imply that a decline in the labor force would not affect agricultural output.

We might touch incidentally upon one particular critique of Schultz's test [159] on the ground that the effect of influenza in causing a decline in the labor force is indiscriminate whereas, with a selective withdrawal of the labor force (from farms where there is surplus labor presumably), there might have been no effect on agricultural output. This argument does not seem to be valid. The assumption here is that the withdrawal of labor (in *some* fashion) with zero impact on agricultural output is necessary or sufficient evidence that there is zero SMP to labor in the economy. But, in the particular example that we have used in the text, labour withdrawal *cannot* but help reduce output even though there is zero shadow wage for labor: and it will make no difference to this qualitative proposition whether the reduction in the population (labor) occurs in the capitalist or the peasant families. Conversely, the fact that labor

who ranks (with Arthur Lewis and Ragnar Nurkse) as the early proponent of the notion of surplus labor. The procedure involved is to take detailed surveys of agricultural output and occupations at the village level, with a view to finding out whether, with unchanged agricultural techniques, and taking full account of the seasonal peaks in demand for labor during harvesting seasons, there exists an excess of labor availability over labor requirements.

This kind of exercise, in essentially the fashion described here, has been deployed in the Indian context by Bhattacharjee [15] for the State of Bihar, during 1957–58 in the course of comprehensive farm management investigations. His estimate of "surplus" labor, fully allowing for the seasonal demand for labor at peak level, runs up to 8.6 per cent of the labor force for North Bihar and 19.8 percent for South Bihar, on application of the Rodan method to male labor alone.

The advantage of this method over the others (insofar as it takes into consideration all opportunities for raising output) is that it goes *directly* to the relevant question: namely, whether there is too much labor in relation to existing availability of techniques (which is what zero SMP or shadow wage of labor means). The real difficulty with the method, on the other hand, is in identifying labor requirements and (in particular) labor availability. The determination of labor availability raises the tricky question of how many hours of work should be fed into the exercise: this is not an easily identifiable technological datum. And the problem could become empirically intractable if we introduced the notion of elastic supply of labor services with respect to rewards.[121] In practice, the estimates have involved adjusting for holidays, festivals, environmental constraints (e.g. "in the month of May the extreme heat makes it physically impossible for any worker to work more than six hours per day in the field" [15]), and then estimating an approximate number of hours which may be expected to be "normal" as far as work is concerned. Shakuntala Mehra [112] has adopted essentially the same approach, in making her estimates of surplus labor in India (from data on labor utilization for 1956–57 and on labor availability for 1961): the only difference consists in explicitly taking her "normal" hours from the "large, capitalist" farms where such normalcy is assumed to obtain (on the ground that surplus labor, and resulting work-sharing, would arise only on the small farms without hired labor). Her estimates also point to the existence of significant amounts of surplus labor in different States (with the exceptions of Gujarat, Maharashtra and Andhra Pradesh). We might note, however, that her actual method is likely to understate the amount of surplus labor, in relation to Bhattacharjee's [15] application of the

can be reduced on some farms without affecting their output may merely reflect the fact that they are totally isolated from the rest of the economy with own factors of production which are not mobile, and the shadow wage of labor for the sector as a whole may be positive (implying "absence of surplus labor").

We may finally note, in reference to the Schultz test, that on examination of the labor force and agricultural output data for the period of the Second World War, for districts in Panjab where there was a significant military draft resulting in declines in the labor force, there appears to have been (according to unpublished work of Manmohan Singh) no impact on the agricultural production in these districts. Clearly, a careful study of this experience (adjusting systematically for trend growth of output resulting from mechanization and impact of possible improvements in the agricultural terms of trade during the War) would be more fruitful than of the influenza episode with its possible complications from factors such as the impact of the epidemic on efficiency.

[121] In this context, we may note that Rosenstein-Rodan and his followers have always noted explicitly that, if those "left behind" did not in fact work the number of hours postulated, in view of "preference for leisure" or on "status" grounds, the actual removal of the estimated surplus labor would reduce output. The recent explorations of the preference-for-leisure hypothesis [158] were thus clearly foreseen by these economists.

Rosenstein–Rodan method at village level, because she estimates the amount of surplus labor on the farms alone: insofar as the landless labor also work less than the postulated "normal" hours, *their* "surplus" labor would be missed out.

Having thus reviewed the literature relating to the major issues in Indian agricultural policy, we now proceed to the final section of our Survey, addressing ourselves to the foreign trade sector.

III. *Foreign Trade*

The Indian policy literature with respect to the foreign sector has been concerned primarily with issues raised by foreign aid, private foreign investment, and trade and exchange rate policies.

(A) *Foreign Aid*

We have already discussed, in Part I, the major issues raised in the Indian literature, relating to the implications of foreign aid for planning investment allocations. The use of "aid to end aid" by a specified time horizon has been the framework within which some important planning exercises have been cast. The political counterpart to such economic analysis has been the appealing notion of ultimate "self reliance"; its conflict with the view that foreign aid must continue as long as the income gap between the affluent and the underdeveloped countries is not drastically reduced has not been noticed. At the same time, economists such as Sengupta [160] have plausibly argued that there is little evidence that the Indian planners have taken seriously their continually receding terminal dates for the net inflow of aid to cease. Both the savings and the trade implications of such a program have been shown by Sengupta to be unrealistic.

The Indian literature has also been concerned with the question of aid tying, by project, by commodity and by source. The prevalence of excess industrial capacity since the Second Plan has been attributed by many economists, including Reddaway [145], to the fact that the foreign aid to India was excessively tied to projects and thus led to creation of more capacity even when the existing capacity was not fully utilized.[123] On the other hand, since the devaluation of June 1966, several Indian economists have felt that there has been too much nonproject, and too little project, aid: a viewpoint which emphasizes that the efficiency of the aid flow may be jeopardized as soon as the aid is tied, whether to projects or to "maintenance" imports.

The source tying of aid has also been widely considered wasteful, although analytical work on this issue is only recent. While there are as yet no quantitative estimates for India, of the cost of source tying when switching possibilities have been exploited, Lal [89] has recently shown, using data supplied by the Imperial Chemical Industries, that these costs may be quite significant in the chemicals sector. Moreover, at an analytical level, Bhagwati [13] has argued that (i) measuring the costs of aid tying by source via estimation of the excess cost of the actual bundle purchased may under–or over–estimate the "true" cost in the Hicksian sense of compensating variation; and (ii) a sharp distinction needs to be drawn between the observed costs and the *minimum* costs that would have been incurred if the recipient country were to exploit fully its switching possibilities (as, in practice, it rarely does). Thus, for example, with reference to the latter point, Bhagwati has argued that India's import licensing system, which specifies items on licenses by source and then makes these licenses totally nontransferable in all respects, results in *double* tying (by source *and* specification) even

[123] On the other hand, other economists such as Bhagwati and Padma Desai [14] have pointed to other, domestic policy induced factors which may have also accounted for such excess capacity. We discuss these factors later, when we survey the trade policy literature.

when the donor country does not itself insist on commodity specification; and that such double tying increases the possible monopolistic charging of prices on aid financed goods beyond what mere source tying might have brought about.[123] Furthermore, both Bhagwati and Honavar [69], who draws upon Indian experience, have highlighted important respects in which the costs of aid tying by source may be understated by such excess cost estimates: distortions of priorities owing to nonavailability of priority items from a tied source when all switching possibilities are exhausted; recurrent excess costs on maintenance, spares and inputs; social waste inherent in techniques unsuitable to local conditions, and other similar factors.[124]

The tying of aid by commodity, essentially P.L. 480 imports, has also attracted considerable controversy. We have already discussed the literature which is concerned with the impact of P.L. 480 wheat imports upon distribution, as also production. We may however observe at this stage (i) that while P.L. 480 imports certainly reduced wheat prices in particular, and foodgrain prices in general, below what they would otherwise have been, *ceteris paribus*, and (ii) that wheat production, being generally responsive to price change, must have been therefore below what it would otherwise have been, *ceteris paribus*, it would be a nonsequitur to argue that therefore imports of wheat under the P.L. 480 program were "undesirable." This question cannot be assessed unless a framework has been devised to examine the optimal prices and quantities of agricultural and foodgrains outputs, in the light of the international and domestic possibilities (including aid availability). In any case, the indictment levelled at P.L. 480 imports by many Indian economists appears to have been, not that aid was tied to these commodities *beyond* what Indians wanted (a possibility that has not been fully investigated in the literature[125]), but that the Indian government itself was keen to get the P.L. 480 aid and that the availability of such aid was determental to the economic interest of the country.[126] Aside from the depressing effect on the resource allocation to agriculture, which we have already touched upon, economists critical of the P.L. 480 program have alleged that the sheer availability of such aid has prevented the government from pushing ahead on the agricultural front organizationally. Insofar as it can be shown that mere drive and organizational energy could have increased agricultural productivity, and that this opportunity–costless gain to the economy

[123] Bhagwati has shown how *both* priorities in respect of what commodity imports should be allowed *and* satisfaction of source constraints could be achieved, without the ill effects of double tying, merely by making the licenses swappable for imports of the specified items from different sources: $100 worth of U.K. lathes, for example, being turned into $100 worth of French lathes, and $100 worth of the French diesel engines into $100 worth of U.K. diesel engines.

[124] Honavar in particular has noted how, in consequence of source tying by most donors, many Indian factories look like "international exhibitions" of machinery from different parts of the world. He emphasizes the fact that such a building up of plants will frequently add to maintenance and inventory costs by eliminating possible economies of scale which follow from plants put together more homogeneously.

[125] The fact that the Indian government was often keen to sign the P.L. 480 agreements does not rule out the possibility that one, and perhaps a principal, motivation (in some years at any rate) might have been a recognition of the possibility that it was easier to get P.L. 480 aid rather than other forms of aid.

[126] Political objections to reliance on P.L. 480 aid have also been numerous. It is, for example, widely felt that the country must feed itself from domestic production. "Food and freedom" is the title of a reputable economist's work on the problem; and it has been argued, again by an academic economist, that genuine independence is impossible if one eats foreign foodgrains! A more respectable political objection has been raised in relation to the use of counterpart rupee funds: it is felt that the availability of such funds, and the desire not to see them accumulate too rapidly, makes it possible for the donor country to incur expenditure within India which would be ruled out for other, ideologically–oriented donor countries. This is believed by some critics to be the case, for example, with respect to expenditure on Indian education out of P.L. 480 rupee proceeds.

was lost, thanks to the P.L. 480 availability, this would be a valid criticism indeed of that aid program—or of *any* aid program which permitted, either directly or through switching, foodgrains to be imported readily. However, such a case is empirically difficult to establish and has not been persuasively made so far despite its plausibility for many economists in the country.[127]

(B) *Private Foreign Investment*

Although private foreign investment in India, whether gross or net of the outflow of (mainly) the pre-Independence British investments, has been relatively unimportant in relation to the official capital transfers, it has attracted considerable attention from the economists. There are basically two types of questions that have been asked: (1) is private equity investment superior to official loan transfers; and (2) what restrictions must be placed on the inflow of private capital, from the viewpoint of social welfare?

(1) The first question is somewhat academic in view of the fact that private investment and official transfers have hardly ever been substitutes in Indian planning: official transfers have nearly always been accepted to the full amount offered and private foreign investment has always fallen short of levels projected in the official documents. Any choice between them is therefore unreal. Nonetheless, Raj [135] has raised this question at an academic level in the Indian context, prompted by the attempt of the Finance Ministry and private industrial interests to liberalize the rules on private foreign investment on strength of the argument, among others, that private equity investment is less expensive than official loan capital. The rather strange proposition which Raj has criticized involves asserting that "since in the case of loan capital . . . both the principal and interest have to be paid over a defined period it is more economic, from the point of view of saving foreign exchange, to depend on foreign equity capital from private sources since only remittances of profits have to be met in foreign exchange and these too will become large only after the enterprises concerned have matured and begun to yield large profits." [135, pp. 21–22]. Raj has countered this view by examining the actual rates of return on equity capital which are available for India and elsewhere, against the average terms of official aid. In any case, it does not seem sensible to argue that aid terms which conceal varying amounts of real transfers of resources should work out in general to be less attractive than commercial terms (or equity capital) if one evaluates the alternatives in terms of an objective function other than the unacceptable one of reducing the short-run outflows of interest and amortisation.

(2) The question of the restrictions on private foreign investment has raised at least two issues of wider interest.

(a) Should there be any restrictions on the *areas* which private investment could enter? While such restrictions have often been urged on political grounds, several economists have also sought them for economic reasons. The notion that private foreign investment should be confined to only the "priority" areas has been widespread in policy discussions. However, this view must be qualified in three ways. (i) Insofar as the investment consists in buying up *existing* capital stock, even in nonpriority areas, the inflow of foreign exchange can always be utilized for "priority" uses. (ii) If the investment involves fresh creation of capacity in "low-priority" sectors, again it must be remembered that if the overall Plan allows for the creation

[127] A notable sceptic of this argument, and the general thesis against the advisability of P.L. 480 imports, is Dantwala [34]. On the other hand, a different, though not overly critical view, is presented by Streeten and Hill [166].

of such capacity anyway, it does not matter whether foreign or domestic investment goes into it. (iii) Where, however, the foreign investment is being offered for areas which are "nonpriority" and hence ruled out from domestic production and availability, there is a real dilemma which cannot be resolved unless again the economist is prepared to estimate the cost (if any) of foregoing the act of foreign investment—assuming that the alternative is the loss of this capital inflow—and ask the planner or the politician whether the presumed noneconomic advantages from ruling out such commodities from domestic production or availability outweigh these economic costs.[128]

On the other hand, some important factors in the Indian context have made governmental restrictions on the entry of foreign capital into specific areas necessary. (i) Since, as we shall presently see, the Indian trade regime has worked on the principle of automatic grant of protection to domestic industries, combined with restrictions on domestic entry operated through industrial licensing, monopoly rents accrue to investments in several activities. Hence, there exists a second best case for regulating entry into areas where the monopoly rents are likely to make the returns to foreign capital exceed its social marginal product.[129] (ii) Furthermore, in a system reliant on foreign aid for maintenance imports, significant linkages can exist between the level of aid inflow and the level of inflow of private capital.[130] Foreign investors often become powerful pressure groups for increasing aid for maintenance imports to keep their capacities better utilized and hence their investments more profitable. In Indian experience, aid loans have thus been secured from donor countries, with commodity specification *combined with provisions for allocations to the firms from these donor countries*. In such a case, provided such provisions are effective despite switching possibilities, there may exist again a case for ensuring that private foreign capital flows into "priority" areas so that the attendant, discriminatory aid allocation is biased towards, rather than against, the priority sectors. Thus, there are *both* domestic and foreign policy distortions which may make regulation of the sectoral *composition* of the private capital inflow desirable. (iii) Yet another argument which has come up in Indian discussions, for regulating the inflow of capital into certain sectors, follows from the fact that the foreign investing interests in some sectors are monopolistic, as in oil, and governmental intervention may help to increase the net payoff accruing to the country from the proposed act of investment. This is a case where governmental intervention becomes necessary, not because of policy induced distortions, but owing to the presence of endogenous distortions (such as the fortuitous presence of monopoly power). How far such governmental regulation is likely to help is of course an issue on which one might be sceptical; this is part of the more general problem arising when, as Dudley Seers has shrewdly put it, "small countries" face "big companies."

[128] Cf. Bhagwati and Desai [14] on these and the other issues we review in the text.

[129] In relation to this question of domestically created monopoly, Bhagwati [11] has also argued that, where components are sold for assembly, the monopoly profits may be made by "overpricing" the components along with raising the product price to a monopolistically-profit-maximizing level. This has the dual advantage of making the latter price look "reasonable" (since the costs can be shown to be higher this way) and also masking the repatriation of profits (which otherwise attract hostile attention). Another aspect of such a phenomenon is that [12] [14] the economist may then observe "value subtracted" or negative value added at international prices, such an observation implying then, not that the process is not worthwhile *in itself*, but that its possible contribution to national income is outweighed by the monopolistic "exploitation" by the investor.

[130] Hence, private capital inflow may have an externality effect in the form of additional aid flow, thus increasing the optimal level at which private capital would be useful to have.

(b) Another area in which governmental regulation has been proposed by several Indian economists relates to the occasional imposition by foreign investors, on their local counterparts, of a contractual prohibition of export to third markets. Kidron [77] has perceptively noted that the bulk of the new quantitative-restrictions-jumping foreign investment in India is by firms who wish to retain their Indian sales *without jeopardising their third country exports.* This is also the case with firms which are basically selling *both* technology and product, who while selling knowhow to India wish at the same time to safeguard their export of products to other markets. In either case, the effect is to interfere with India's export potential, particularly as India is increasingly relying at the margin on the exports of her newer manufactures. Economists such as Raj have therefore pressed for the prohibition of clauses restricting exports from India.

(C) *Trade and Exchange Rate Policies*

The literature on India's trade and exchange rate policies, involving the entire effective exchange rate system, has also raised some issues of general interest. Among other things, it has cast additional light on the drawbacks of a regime involving continued reliance on quantitative restrictions (especially when operated so as to provide automatic protection) and a pattern of reluctant exchange rate adjustments.

Import Controls: Beginning essentially with the 1956–57 foreign exchange crisis, India has been on a strict import and exchange control system. Furthermore, this system has been administratively operated (at least until the June 1966 devaluation of the Indian rupee) such that (1) all industrial capacity creation has been regulated by industrial licensing, extending to the so-called CG licensing of imported capital goods and (2) most input and raw material allocations have been allocated via the so-called "actual user" (AU) licenses, directly to producers.[131]

The allocations under the AU category, to which in particular considerable attention has been directed by economists [162] [14], have been worked on two basic principles: (i) "essentiality" and (ii) indigenous nonavailability". For every AU import, some specified agencies of the government must certify that they are "essential" for production *and* that they are not available from domestic sources. Since the latter principle has been operated virtually without reference to the cost of domestic production, it has amounted to giving automatic and anticipatory" protection to domestic industries. Panchmukhi, Bhagwati and Padma Desai [126], who have estimated the resulting effective rates of protection, for different Indian industrial processes for 1961 and 1962, have found these protective rates going up to levels as high as over 10,000 per cent, with others at almost as high negative values: the range thus being enormous. These results merely underline the totally unpredictable, extreme and often bizarre nature of protection given by a QR-regime operated on the principle of automatic protection.[132] In this connection, the devalued role of the Tariff Commission, whose work has been studied in depth by Padma Desai [44], and the critique of the new era of indiscriminate protection through QR-policy in his Presidential Address to the Indian Economic Association by Lakdawala [87], are of some interest.

Bhagwati and Desai [14] have further noted that the principle of automatic protection issuing from QR's creates a bias in

[131] In addition, of course, there have been licenses for export promotion under the import entitlement schemes, and other minor categories. For detailed description, see Shourie [162] as also Bhagwati and Desai [14].

[132] Panchmukhi, Bhagwati and Desai [126] also raise some conceptual questions of importance relating to the notion of effective protection when the calculations are based on import-premia-determined implicit rates of tariff.

favor, *ceteris paribus*, of industries with imported, as distinct from domestically produced, inputs. Insofar as the *quantity* of import allocations tends to be inversely related to the availability of indigenously produced inputs, under such a system, there would result a bias in the effective incentive provided to the processes using relatively more imported inputs: they would be able to get relatively greater allocations of imports under AU licenses and hence obtain these inputs at import–premium–*exclusive* prices (which would include only the explicit tariff duty) whereas the other industries would have to buy import substitute, indigenous items at premium *inclusive* prices (since these items would fetch a price equal to the c.i.f. price plus the import premium). The effective incentive given to the former industries or processes would thus be greater, *ceteris paribus*.

Furthermore, aside from the traditional discussion of delays, lack of coordination among different licensing agencies and similar administrative deficiencies which reduce the efficiency of a QR–regime, the Indian import control policy has also been alleged to have operated, in the ultimate analysis, without any *economic* criteria [14] [162].[133] Economists investigating these criteria have argued that these are rarely defined; that (in view of the multitude of activities which demand these import allocations) they could hardly be defined; and that in practice rules of thumb have had to be used, these rules often (though not always) taking some notion of "equitable" distribution as the operative guiding principle. This finding has its counterpart in the conclusions of the Raj Committee's Report of Steel Control [178], and indeed in nearly all the empirical studies relating to the working of the controls of scarce materials.[134]

Concerning the particularly widespread rule of thumb which related the AU allocations of materials to installed capacity, it has further been argued [14] that this procedure creates a bias towards the creation of capacity despite the underutilization of existing capacity. This may be because an entrepreneur who wishes to extend capacity utilization may not be able to do so as legal access to more materials is virtually ruled out by the import licensing system (except since recently, on a limited account, through the import entitlement licenses marketed by exporter). However, even if access to such materials were freely available, the fact that additional capacity installation would result in *pro rata* grant of import–premium–exclusive imports under AU licenses whereas additional utilization of existing capacity must be through purchase of import–premium–inclusive materials from the market, would bias the choice at the margin in favor of the former course.

Moreover, the fact that the reliance on QR's also implies a loss of "revenue," in relation to an import rate change which would mop up the premium, has been among the principal motivating factors behind the Indian literature proposing an exchange auction system, which was sug-

[133] Whether even the Mahalanobis–type strategy has been consistently followed in this area has been a matter of some controversy. Hazari [68] has argued for instance that Indian luxury goods production (and hence consumption) has been allowed to absorb, directly and indirectly, a fraction of the available foreign exchange which is not negligible. Of course, whether the luxury expenditure would not otherwise have been diverted into other areas where it might have cut into exports, for example, needs to be investigated in order to arrive at a more adequate picture.

[134] In this general connection, the following quote from Raj Committee [178] is particularly revealing: "As regards priorities, the Iron and Steel Controller gives different ratings of priority according to the nature of each case. Thus some indents receive 'over-riding' priority and others 'top priority'; the categories of priority have further proliferated and we understand that there is now even a category of 'red hot priority'! As already pointed out, the Iron and Steel Controller's Office does not have with it data relating to outstanding orders with the producers classified according to priority and non–priority indents . . . there is no systematic checking as to whether the priorities are in fact being respected by the producers."

gested by Bhagwati [5] after examining its compatibility with Indian planning objectives, and the alternative proposal to use tariffs more freely for this purpose, as indeed Indian budgets have recently been designed to do.

Export Policy: However, these proposals will not directly moderate, or eliminate, the disincentive against exports that an over-valued exchange rate constitutes. The governmental measures aimed at eliminating this bias against exports have taken the form principally of import entitlement schemes (under which premium earning import licenses are given to eligible exporters on a *pro rata* basis related to f.o.b. export values). These schemes have come in for scrutiny from economists such as Gulati [60] and Bhagwati [12]. The ad hoc manner in which the resulting export incentives were granted to a whole range of industries has been argued to have provided a parallel to the "indiscriminate" protection from imports conferred on domestic production by the QR's. The allocative inefficiency of these schemes as export promotion measures has been underlined by pointing to the extreme phenomenon of negative value added (at international prices) which can and did arise in the Indian context, thanks to the difference between the f.o.b. value of exports and the c.i.f. value of imports (and import substitutes) being made up by export subsidies [12].[135]

Exchange Rate Policy: An appreciation of these and other inefficiencies underlying the governmental policies designed to simulate, but avoid, a formal devaluation via export subsidies and tariffs has prompted some Indian economists to press for formal parity changes. While the economists' attitudes towards devaluation have undergone a change on the Indian scene, simultaneously with a greater appreciation of the role of the price mechanism even in a socialist framework,[136] there has been a certain degree of scepticism about the timing of the devaluation in June 1966 and a much greater degree of political opposition in view of the fact that the pressure for it was brought largely by the donor countries through the I.B.R.D. Aid Consortium (even though all influential economists within the country were by no means opposed to the measure). The second, severe agricultural drought which overlaid the devaluation must have led to a significant rise in the domestic prices of agricultural and agriculture-based items, which in turn impeded any significant improvement in exports, thus creating the impression that the devaluation had *caused* the rise in prices and had *failed* because exports continued to stagnate. In point of fact, it is arguable that many Indian exports would have been priced out internationally were it not for the devaluation, as the price rise was largely thanks to the second drought and thus autonomous of the exchange rate change.[137] Unfortunately, no serious empirical analysis of this important policy decision has yet been forthcoming, the field having been left in the popular debate to economists whose analysis leaves much to be desired[138] but

[135] An extended evaluation of the entitlement schemes from the viewpoint of their *economic* efficiency is contained in Bhagwati and Desai [14].

[136] Raj's [136] thoughtful piece on this general problem makes interesting reading. The controversy in 1962 between Bhagwati [7], who argued for a freer use of exchange rate changes in the shape of a devaluation, and Bardhan [1] and Dasgupta [38] who argued the opposite case, is also of some interest in this general context. Bardhan's general position that a devaluation can be simulated by equivalent import duties and export subsidies, and hence is not necessary, has been later discussed in Bhagwati [12].

[137] Further, the reduction of overinvoicing of exports of the newer manufactures, whose export subsidies were withdrawn with the devaluation (as part of the general rationalization, the *net* devaluation being therefore much less than the apparent one), must also be considered. The devaluation must also be judged as a measure of rationalization, the older methods of *de facto* devaluation having been largely scrapped with its introduction.

[138] Thus, for example, it is not uncommon to compare the twelve-monthly returns *immediately* after the June

whose critical views have been expressed with considerable conviction.

IV. *Concluding Remarks*

In conclusion, it is perhaps worth emphasizing the selective nature of our Survey. We expect, however, to have reviewed much of the policy literature with an analytical base and related literature that has developed against the backdrop of policy issues.[189]

It is clear that the Survey highlights both the similarity of the Indian analyses of policy issues with that in many other developing countries, as also some striking differences endemic to the Indian economy and scene. On the one hand, we have noted the concern of Indian economists with familiar issues such as trade and exchange rate policies, foreign aid and private foreign capital, and response of agricultural production to price change. On the other hand, the *structural* planning models, the analysis of choice-of-technique problems on the assumption of a *labor surplus* economy and the debate on foodgrains policy in terms of *zonal restrictions* (emphasizing India's federal setup) underline the somewhat uncommon character of India's economy and political structure.

---

devaluation with the preceding twelve months, with no awareness of lags. Nor is any notice taken of the factors mentioned in the text and in the previous footnote. Nor is there any awareness of the fact that, at least for three months subsequent to the devaluation and the attendant elimination of the import entitlement schemes for exporters, the Minister for International Trade was keen to restore these schemes and went around saying that he would succeed in doing so: the effect of this on the export performance in the period after the devaluation should not be ignored.

[189] Among the issues which have been prominent, but which we have decided to omit from the Survey, are (1) whether income and wealth inequalities have been accentuated during the three Plans, with related questions about the trends in the real income of *agricultural* landless labor and in the concentration of *industrial* capacity and invested capital in the hands of a few top "industrial houses," and (2) whether decentralization in rural administration and planning, via the so-called *Panchayati Raj* system, has been beneficial for agricutlural planning and growth [62].

REFERENCES

1. P. K. BARDHAN, "The Case Against Devaluation," *The Economic Weekly*, Sept. 1962, *14*, 1448.
2. F. M. BATOR, "On Capital Productivity, Input Allocation and Growth," *Quart. Jour. Econ.*, Feb. 1957, *71*, 86–106.
3. C. BETTELHEIM, "On the Problem of Choice Between Alternative Investment Projects," *Soviet Studies*, July 1950, *2*, 22–42.
4. J. BHAGWATI, "Unemployment, Relative Prices and the Saving Potential," *The Indian Economic Review*, Aug. 1958, *4*, 71–82.
5. ———, "Indian Balance of Payments Policy and Exchange Auctions," *Oxford Econ. Papers*, Feb. 1962, *14*, 51–68.
6. ———, "On How to Decide What to Import and What to Produce," *Oxford Econ. Papers*, Feb. 1962, *14*, 114–116.
7. ———, "The Case for Devaluation," *The Economic Weekly*, Aug. 1962, *14*, 1263–1266.
8. ———, "The Theory of Comparative Advantage in the Context of Underdevelopment and Growth," *The Pakistan Development Review*, Autumn 1962, *2*, 339–353.
9. ———, *The Economics of Underdeveloped Countries*. London 1966.
10. ———, "Review of A. H. Hanson: The Process of Planning: A Study of India's Five Year Plans, 1950–1964," *Public Administration*, Summer 1967, *45*, 214–217.
11. ———, "Fiscal Policies, The Faking of Foreign Trade Declarations, and the Balance of Payments," *Bulletin of the Oxford University Institute of Economics and Statistics*, 1967, *29*, 61–77.
12. ———, *Theory and Practice of Commercial Policy: Departures from Unified Exchange Rates*. Princeton 1968.
13. ———, "The Tying of Aid," *UNCTAD Secretariat* (New York), TD/7/Supp. 4, 1967.
14. J. BHAGWATI AND PADMA DESAI, *Planning for Industrialization: India's Trade and Industrialization Policies (1950–1966)* Oxford 1970, forthcoming.

15. J. P. BHATTACHARJEE, "Underemployment Among Indian Farmers: An Analysis of Its Nature and Extent Based on Data for Bihar," *Artha Vijnana*, Sept. 1961, *3*, 246–279.
16. A. K. BISWAS, "Optimal Technique in Multi-sector Economy," *The Economic Weekly*, Aug. 1957, *9*, 1055–1058.
17. ———, "On a Model of Planned Growth," *Arthaniti*, Nov. 1957, *1*, 40–56.
18. P. R. BRAHMANAND, "The General Theory and the Problem of Development of Backward Areas," *Journal of the University of Bombay*, July 1952, *21*, 53–83.
19. ———, and C. N. Vakil, *Planning for an Expanding Economy*. Bombay 1956.
20. S. CHAKRAVARTY, "The Mahalanobis Model of Development Planning," *Arthaniti*, Nov. 1957, *1*, 57–69.
21. ———, "The Mathematical Framework of the Third Five Year Plan," in P. N. Rosenstein-Rodan, ed., *Capital Formation and Economic Development*, London 1964, pp. 11–22.
22. ———, "Optimal Programmes of Capital Accumulation in a Two-Sector Model with Non-Shiftable Capital," Jan. 1967 (mimeographed).
23. ———, "Alternative Preference Functions in Problems of Investment Planning on the National Level," in E. Malinvaud and M. O. L. Bacharach, ed., *Activity Analysis in the Theory of Growth and Planning*, New York 1967, pp. 150–169.
24. S. CHAKRAVARTY AND R. S. ECKAUS, "An Approach to a Multi-Sectoral Intertemporal Planning Model," in P. N. Rosenstein-Rodan, ed., *Capital Formation and Economic Development*, London 1964, pp. 110–127.
25. S. CHAKRAVARTY AND L. LEFEBER, "An Optimizing Planning Model," *The Economic Weekly*, Feb. 1965, *17*, 237–252.
26. S. CHAKRAVARTY AND P. N. ROSENSTEIN-RODAN, *Linking Food Aid With Other Aid*. Rome 1965.
27. V. M. DANDEKAR, *Use of Food Surpluses for Economic Development*. Poona 1956.
28. ———, "A Review of the Land Reform Studies Sponsored by the Research Programmes Committee of the Planning Commission," *Artha Vijnana*, Dec. 1962, *4*, 291–329.
29. ———, "From Agarian Reorganisation to Land Reform," *Artha Vijnana*, Mar. 1964, *6*, 51–70.
30. ———, "Economic Theory and Agrarian Reform," *Oxford Econ. Papers*, Feb. 1962, *14*, 69–80.
31. ———, "Prices, Production and Marketed Supply of Foodgrains," *Indian Journal of Agricultural Economics*, July–Dec. 1964, *19*, 186–195.
32. ———, "Minimum Support Prices for Foodgrains: Guidelines for a Policy and a Programme," *Artha Vijnana*, Dec. 1965, *7*, 273–284.
33. ———, "Agricultural Price Policy: A Critique of Dantwala," *Economic and Political Weekly*, Mar. 1968, 454–459.
34. M. L. DANTWALA, "Incentives and Disincentives in Indian Agriculture," *Indian Journal of Agricultural Economics*, April–June 1957, *22*, 1–25
35. ———, "Impact of Redistribution and Pooling of Land on Agrarian Structure and Efficiency of Resource Use," *Indian Journal of Agricultural Economics*, Dec. 1959, *14*, 80–85.
36. ———, "Growing Irrelevance of Economics in Planning: Case of Procurement Prices," *Economic and Political Weekly*, Oct. 1967, 1945–1947.
37. ———, "Agricultural Price Policy: Reply," *Economic and Political Weekly*, Mar. 1968, 459–461.
38. A. K. DAS GUPTA, "Export Promotion Through Devaluation," *The Economic Weekly*, Sept. 1962, *14*, 1146–1447.
39. M. DATTA-CHAUDHURI, "Regional Specialisation in Metallurgical and Machine-Building Industries in India in the Framework of a Planning Model for Optimum Use of National Resources," *The Indian Economic Review*, Oct. 1967, *2* (New Series), 219–270.
40. D. K. DESAI, *Increasing Income and*

Production in Indian Farming. Bombay 1963.

41. P. DESAI, "A Short-term Planning Model for the Indian Economy," *Rev. Econ. Stat.*, May 1961, *43*, 193–200.
42. ———, "The Development of the Indian Economy: An Exercise in Economic Planning," *Oxford Econ. Papers*, Nov. 1963, *15*, 308–317.
43. ———, "Fiscal and Monetary Equilibrium in the Formulation of National Plans," *Economia Internazionale*, Aug. 1968, *21*, 3–15.
44. ———, *Tariff Protection in India*, Delhi School of Economics 1969.
45. P. N. DHAR AND H. F. LYDALL, *The Role of Small Enterprises in Indian Economic Development*. Delhi 1961.
46. A. K. DIXIT, "Optimal Development in the Labour Surplus Economy," *The Rev. Econ. Stud.*, Jan 1968., *35*, 23–34.
47. E. D. DOMAR, "A Soviet Model of Growth," in *Essays in the Theory of Growth*, New York 1957, pp. 223–261.
48. R. C. DUTT, *Economic History of India in the Victorian Age*. London 1956.
49. R. S. ECKAUS, "Planning in India," in M. F. Millikan, ed., *National Economic Planning*, New York 1967, pp. 305–378.
50. R. S. ECKAUS AND L. LEFEBER, "Capital Formation: A Theoretical and Empirical Analysis," in P. N. Rosenstein–Rodan, ed., *Capital Formation and Economic Development*, London 1964, pp. 83–109.
51. R. S. ECKAUS AND K. S. PARIKH, *Planning for Growth*. Cambridge, Mass. 1968.
52. T. S. EPSTEIN, *Economic Development and Social Change in South India*. Bombay 1962.
53. R. FRISCH, "Oslo Decision Models," 1957 (mimeographed).
54. ———, *Planning for India: Selected Explorations in Methodology*. Calcutta 1960.
55. D. R. GADGIL, *Planning and Economic Policy in India*. Poona 1961.
56. V. GANDHI, *Tax Burden on Indian Agriculture*. Unpublished Doctoral Dissertation, Harvard University, May 1964.
57. N. GEORGESCU–ROEGEN, "Economic Theory and Agrarian Economics," *Oxford Econ. Papers*, Feb. 1960, *12*, 1–40.
58. A. Ghosh, "Mathematical Models in Research Related to Planning in India: A Review," *Economics of Planning*, 1966, *6*, 154–178.
59. D. GHOSH, "Technique of Production and Employment in an Underdeveloped Economy," *The Economic Weekly*, Aug. 1956, *8*, 919–920.
60. I. S. GULATI, "Export Promotion Through Import Entitlement," *The Economic Weekly*, May 1965, *17*, 859–864.
61. S. C. GUPTA, "Agricultural Economics in India," *International Journal of Agrarian Affairs*, Dec. 1966, *5*, 1–51.
62. A. H. HANSON, *The Process of Planning: A Study of India's Five Year Plans 1950–1964*. London 1966.
63. C. H. HANUMANTHA RAO, "Farm Size and Economies of Scale," *The Economic Weekly*, Dec. 1963, *15*, 2041–2044.
64. ———, *Agricultural Production Functions, Costs and Returns in India*. Bombay 1965.
65. ———, "The Marketable Surplus Function for a Subsistence Crop: Comments," *The Economic Weekly*, Apr. 1965, *17*, 677–678.
66. ———, "Alternative Explanations of the Inverse Relationship Between Farm Size and Output per Acre in India," *The Indian Economic Review*, Oct. 1966 *1* (New Series), 1–12.
67. ———, "Entrepreneurship, Management and Farm Tenure Systems," Feb. 1967, mimeographed.
68. B. R. HAZARI, "Import-Intensity of Consumption in India," *The Indian Economic Review*, Oct. 1967, *2* (New Series), 155–176.
69. R. M. HONAVAR, "Aid–Tying by Source and Industrial Efficiency," *Economic Development Institute* (I.B.R.D.), mimeographed, 1967.
70. W. D. HOPPER, "Allocation Efficiency

in a Traditional Indian Agriculture," *Jour. Farm Econ.*, Aug. 1965, *47*, 611–624.

71. D. R. KHATKHATE, "Some Notes on the Real Effects of Foreign Surplus Disposal in Underdeveloped Economies," *Quart. Jour. Econ.*, May 1962, *76*, 186–196.
72. A. M. KHUSRO, *Economic and Social Effects of Jagirdari Abolition and Land Reforms in Hyderabad*. Planning Commission (Research Programmes Committee), Unpublished 1960.
73. ———, *Economic Development with No Population Transfers*. New Delhi 1962.
74. ———, "Returns to Scale in Indian Agriculture," *Indian Journal of Agricultural Economics*, Oct.–Dec. 1964, *19*, 51–80.
75. ———, "The Pricing of Food in India," *Quart. Jour. Econ.*, May 1967, *81*, 271–285.
76. ———, "Economic Theory and Indian Agricultural Policy," in A. M. Khurso, ed., *Readings in Agricultural Development*, Delhi 1968.
77. M. KIDRON, *Foreign Investments in India*. London 1965.
78. R. KOMIYA, "A Note on Professor Mahalanobis' Model of Indian Economic Planning," Feb. 1959, *Rev. Econ. Stat.*, *41*, 29–35.
79. J. KRISHNA AND M. S. RAO, "Price Expectation and Acreage Response for Wheat in Uttar Pradesh," *Indian Journal of Agricultural Economics*, Jan.–Mar. 1965, *20*, 20–25.
80. R. KRISHNA, "A Note on the Elasticity of the Marketable Surplus," *Indian Journal of Agricultural Economics*, July–Sept. 1962, *17*, 79–84.
81. ———, "Farm Supply Response in India–Pakistan: A Case Study of the Panjab Region," *Econ. Jour.*, Sept. 1963, *73*, 477–487.
82. ———, "The Marketable Surplus Function for a Subsistence Crop: An Analysis with Indian Data," *The Economic Weekly*, Feb. 1965, *17*, 309–320.
83. ———, "The Marketable Surplus Function for a Subsistence Crop: Reply to Comments," *The Economic Weekly*, Nov. 1965, *17*, 1665–1668.
84. ———, "Government Operations in Foodgrains," *Economic and Political Weekly*, Sept. 1967, 1695–1706.
85. K. KRISHNAMURTY AND D. U. SHASTRI, "Inventories in the Organised Manufacturing Sector of the Indian Economy: A Disaggregative Study," 1966, mimeographed.
86. T. N. KRISHNAN, "The Marketable Surplus of Foodgrains: Is It Inversely Related to Price?", *The Economic Weekly*, Feb. 1965, *17*, 325–328.
87. D. T. LAKDAWALA, "Aspects of Trade Policy in India," *Indian Economic Journal*, Oct.–Dec. 1964 *12*, 89–110.
88. D. T. LAKDAWALA AND J. C. SANDESARA, *Small Industry in a Big City: A Survey in Bombay*. Bombay 1960.
89. D. LAL, "Cost of Aid–Tying: A Case Study of India's Chemical Industry," *UNCTAD Secretariat* (New York), 1967.
90. U. LELE, "Market Integration: A Study of Sorghum Prices in Western India," *Jour. Farm Econ.*, Feb. 1967, *49*, 147–159.
91. M. LIPTON, "Should Reasonable Farmers Respond to Price Changes?", *Modern Asian Studies*, Jan. 1967, *1*, 95–99.
92. I. M. D. LITTLE, "A Critical Examination of India's Third Five Year Plan," *Oxford Econ. Papers*, Feb. 1962, *14*, 1–24.
93. ———, "Real Cost of Labour, and the Choice Between Consumption and Investment," in P. N. Rosenstein–Rodan, ed., *Pricing and Fiscal Policies*, London 1964, pp. 77–91.
94. ———, "Tax Policy and the Third Plan," in P. N. Rosenstein–Rodan, ed., *Pricing and Fiscal Policies*, London, 1964, pp. 30–76.
95. P. C. MAHALANOBIS, "Some Observations on the Process of Growth of National Income," *Sankhya*, Sept. 1953, *12*, 307–312.
96. ———, "The Approach of Operational Research to Planning in India," *Sankhya*, Dec. 1955, *16*, 3–130.

97. I. M. MAJUMDAR, "Marketable Surplus Function for a Subsistence Crop: Further Comments," *The Economic Weekly*, May 1965, *17*, 820–822.
98. J. S. MANN, "The Impact of P.L. 480 Imports on Prices and Domestic Supply of Cereals in India," *Jour. Farm Econ.*, Feb. 1967, *49*, 131–146.
99. A. S. MANNE, ed., *Investments for Capacity Expansion*. London 1967.
100. A. S. MANNE AND J. BERGSMAN, "An Almost Consistent Intertemporal Model for India's Fourth and Fifth Plans," in I. Adelman and E. THORBECKE, ed., *The Theory and Design of Economic Development*, Baltimore 1966, pp. 239–256.
101. A. S. MANNE AND A. RUDRA, "A Consistency Model of India's Fourth Plan," *Sankhya* (Series B), Feb. 1965, *27*, 57–144.
102. A. S. MANNE AND T. E. WEISSKOPF, "A Dynamic Multisectoral Model for India: 1967–1975," Dec. 1967, mimeographed.
103. S. A. MARGLIN, *Approaches to Dynamic Investment Planning*. Amsterdam 1963.
104. ———, "Industrial Development in the Labour-Surplus Economy: An Essay in the Theory of Optimal Growth," Jan. 1966, mimeographed.
105. ———, *Public Investment Criteria: Benefit-Cost Analysis for Planned Growth*. London 1967.
106. E. S. MASON, *Economic Development in India and Pakistan*. Harvard 1966.
107. P. N. MATHUR, "Time Pattern and Quanta of Purchase and Sale of *Jawar* by the Peasants of Vidarbha (Bombay State)," *Artha Vijnana*, 1959, *1*, 61–69.
108. ———, "Differential Effects of Price Increase on Small and Big Cultivators —A Case Study," *Artha Vijnana*, 1962, *4*, 44–51.
109. ———, "An Efficient Path for the Technological Transformation of an Economy," in T. Barna, ed., *Structural Interdependence and Economic Development*, London 1963, pp. 39–56.
110. P. N. MATHUR AND H. EZEKIEL, Marketable Surplus of Food and Price Fluctuations in a Developing Economy," *Kyklos*, 1961, *14*, 396–408.
111. D. MAZUMDAR, "On the Economics of Relative Efficiency of Small Farmers," *The Economic Weekly*, July 1963, *15*, 1259–1263.
112. S. MEHRA, "Surplus Labour in Indian Agriculture," *The Indian Economic Review*, April 1966, *1* (New Series), 111–129.
113. B. S. MINHAS, "Measurement of Agricultural Growth: Rapporteur's Report," *Indian Journal of Agricultural Economics*, Oct.-Dec. 1966, *21*, 165–182.
114. ———, "Water Requirements of Crops and Efficiency of Irrigation Water in India," Indian Statistical Institute (Planning Unit), Discussion Paper No. 24, Aug. 1967, mimeographed.
115. B. S. MINHAS AND T. N. SRINIVASAN, "New Agricultural Strategy Analysed," *Yojana*, Jan. 1966, 20–24.
116. B. S. MINHAS AND A. VAIDYANATHAN, "The Growth of Agricultural Output in India, 1951–54–1958–61: An Analysis of Component Elements," *Journal of the Indian Society of Agricultural Statistics*, December 1965, *17*, 230–252.
117. A. MITRA, "A Note on the Mahalanobis Model," *The Economic Weekly*, Mar. 1957, *9*, 372–374.
118. D. NAOROJI, *Poverty and Un-British Rule in India*. Delhi 1962.
119. D. NARAIN, "Ratio of Interchange Between Agricultural and Manufactured Goods in Relation to Capital Formation in Underdeveloped Economies," *Indian Economic Review*, Aug. 1957, *3*, 46–55.
120. ———, *Distribution of the Marketed Surplus of Agricultural Produce by Size-Level of Holdings in India: 1950–51*, Bombay 1962.
121. ———, *Impact of Price Movements on Areas Under Selected Crops in India: 1900–1939*. Cambridge 1965.
122. N. V. A. NARASIMHAN, *A Short-term Planning Model for India*. Amsterdam 1956.
123. V. F. NOWSHIRVANI, "Allocation Effi-

ciency in a Traditional Indian Agriculture," *Jour. Farm Econ.*, Feb. 1967, *49*, 218–221.
124. ———, "A Note on the Elasticity of the Marketable Surplus—A Comment," *Indian Journal of Agricultural Economics*, Jan.–Mar. 1967, *22*, 110–114.
125. ———, "A Note on the Fixed-Cash Requirement Theory of Marketed Surplus in Subsistence Agriculture," *Kyklos*, 1967, *20*, 772–773.
126. V. R. Panchmukhi, J. Bhagwati and P. Desai, "Effective Rates of Protection in India: 1961 and 1962," Feb. 1968, mimeographed.
127. A. Parikh, "Statewise Growth Rate in Agricultural Output—An Economic Analysis," *Artha Vijnana*, Mar. 1966, *8*, 1–52.
128. R. P. Patwardhan and D. V. Ambekar, ed., *The Speeches and Writings of Gopal Krishna Gokhale. 1*, Bombay 1962.
129. P. H. Prasad, "Sectoral Models of Growth—A Critique," *The Indian Economic Review*, Apr. 1966, *1* (New Series), 104–111.
130. K. N. Raj, "Small-Scale Industries—Problems of Technological Change: Part I," *The Economic Weekly*, Apr. 1956, *8*, 420–421.
131. ———, "Small-Scale Industries—Problems of Technological Change: Part II," *The Economic Weekly*, Apr. 1956, *8*, 435–438.
132. ———, *Employment Aspects of Planning in Underdeveloped Economies.* Cairo 1957.
133. ———, *Some Economic Aspects of the Bhakra-Nangal Project.* Delhi 1960.
134. ———, "Growth Models in Indian Planning," *The Indian Economic Review*, Feb. 1961, *5*, 242–261.
135. ———, *Indian Economic Growth: Performance and Prospects.* Delhi 1965.
136. ———, "What Does Socialism Imply for Economic Policy Now?", *Mainstream*, Jan. 1964, *2*, 19–22.
137. ———, "The Current Economic Crisis," *Times of India*, Feb. 1966.
138. ———, *India, Pakistan and China: Economic Growth and Outlook.* Delhi 1966.
139. ———, "Investment in Livestock in Agrarian Economies: A Theoretical and Empirical Analysis," Dec. 1967, mimeographed.
140. K. N. Raj and A. K. Sen, "Alternative Patterns of Growth Under Conditions of Stagnant Export Earnings," *Oxford Econ. Papers*, Feb. 1961, *13*, 43–52.
141. M. G. Ranade, *Essays on Indian Economics.* Bombay 1898.
142. A. P. Rao, "Size of Holdings and Productivity," *Economic and Political Weekly*, Nov. 1967, 1989–1991.
143. V.K.R.V. Rao, "Investment, Income and Multiplier in an Underdeveloped Economy," *The Indian Economic Review*, Feb. 1952, *1*, 55–67.
144. N. Rath and V. S. Patwardhan, *Impact of Assistance Under P.L. 480 on Indian Economy.* Poona 1967.
145. W. B. Reddaway, *The Development of the Indian Economy*, London 1962.
146. ———, "*The Development of the Indian Economy:* The Objects of the Exercise Restated," *Oxford Econ. Papers*, Nov. 1963, *15*, 318–332.
147. J. Robinson, "The Choice of Techniques," *The Economic Weekly*, June 1956, *8*, 715–718.
148. P.N. Rosenstein-Rodan, "Contribution of Atomic Energy to a Power Programme," in P.N. Rosenstein-Rodan, ed., *Pricing and Fiscal Policies*, London 1964. pp. 165–174.
149. A. Rudra, "The Problem of Technological Choice," *The Economic Weekly*, May 1956, *8*, 556–558.
150. J. Sandee, *A Demonstration Planning Model for India.* Calcutta 1960.
151. T. W. Schultz, *Transforming Traditional Agriculture.* New Haven 1964.
152. ———, "Significance of India's 1918–19 Losses of Agricultural Labour—A Reply," *Econ. Jour.*, Mar. 1967, *77*, 161–163.
153. A. K. Sen, "Labour Cost and Economic Growth," *The Economic Weekly*, September 1956, *8*, 1159–1160.
154. ———, "Unemployment, Relative

Prices and the Saving Potential," *The Indian Economic Review*, Aug. 1957, *3*, 56–63.

155. ———, "A Short Note on the Ambar Charkha," *The Economic Weekly*, Oct. 1957, *9*, 1357–1358.
156. ———, *Choice of Techniques: An Aspect of the Theory of Planned Development*, 2nd ed. Oxford 1962.
157. ———, "Size of Holdings and Productivity," *The Economic Weekly*, Feb. 1964, *16*, 323–326.
158. ———, "Peasants and Dualism with or without Surplus Labour," *Jour. Pol. Econ.*, Oct. 1966, *74*, 425–450.
159. ———, "Surplus Labour in India: A Critique of Schultz's Statistical Test," *Econ. Jour.*, Mar. 1967, *77*, 154–161.
160. A. Sengupta, "Foreign Capital Requirements in India's Perspective Planning," *The Economic Weekly*, Aug. 1965, *17*, 1267–1276.
161. C. H. Shah, "Problems of Small Farmers: A Report on an Enquiry into the Problems of Low Income Farmers in Kodinar Taluka," Unpublished Paper prepared for the Research Programmes Committee, Planning Commission, India. 1965
162. A. Shourie, *Allocation of Foreign Exchange in India*. Unpublished doctoral dissertation, Syracuse University, 1966.
163. M. Singh, *India's Export Trends and the Prospects for Self-Sustained Growth*. Oxford 1964.
164. N. V. Sovani and N. Rath, *Economics of a Multiple-Purpose River Dam*. Poona 1960.
165. T. N. Srinivasan, "Critique of the Optimizing Planning Model," *The Economic Weekly*, Feb. 1965, *17*, 255–264.
166. P. Streeten and R. Hill, "Foreign Aid to India," in P. Streeten and M. Lipton, ed., *The Crisis of Indian Planning*, London 1968.
167. R. Strotz, "Myopia and Inconsistency in Dynamic Utility Maximization," *Rev. Econ. Stud.*, 1956, *23*, 165–180.
168. K. Sundaram, "P.L. 480 Imports: Efficiency in Purchase and Distribution," *Economic and Political Weekly*, Mar. 1967, 487–490.
169. P. Thakurdas (*et. al*), *A Plan of Economic Development for India*. London 1945.
170. J. Tinbergen, *The Design of Development*. Baltimore 1958.
171. ———, "Mill Versus Cottage Industries," *The Economic Weekly*, Jan. 1957, *9*, 81–83.
172. S. Tsuru, "Some Theoretical Doubts on the Plan Frame," *The Economic Weekly*, Jan. 1957, *9*, 77–79.
173. L. S. Venkataramanan, "A Statistical Study of Indian Jute Production and Marketing with Special Reference to Foreign Demand." Unpublished doctoral dissertation, University of Chicago, June 1958.
174. ———, "Hedging Stocks and Futures Contract in India: Cotton Industry," in D. Kumar, S. P. Nag and L. S. Venkataramanan, *Resource Allocation in the Cotton Textile Industry*, Bombay 1965, pp. 135–185.
175. T. E. Weisskopf, "A Programming Model for Import Substitution in India," *Sankhya* (Series B), Dec. 1967, *29*, 257–306.
176. Government of India, Ministry of Agriculture, *Report of the Agricultural Prices Commission on Price Policy for Kharif Cereals for 1965–66*, Delhi 1965.
177. Government of India, Ministry of Food, Agriculture, Community Development and Cooperation, *Report of the Foodgrains Policy Committee*, Delhi 1966.
178. Government of India, Ministry of Steel and Heavy Industries, *Report of Steel Control*, Delhi 1963.
179. Government of India, Planning Commission, *First Five Year Plan*, Delhi 1952.

Economic Policy Discussion and Research in Israel

By NADAV HALEVI*

TABLE OF CONTENTS

Economic Policy Discussion and Research in Israel

By NADAV HALEVI*

The economic policy issues in Israel and the discussions they aroused must be viewed against a background of the economic development of the country. This cannot be surveyed here,[1] but a very brief indication of the salient points and of related policy issues may not be amiss.

After gaining its independence in May 1948, Israel embarked on an economic policy marked by large scale immigration and the need for rapid economic development to absorb the new immigrants. Though Israelies often describe Israel as a country steadily receiving a heavy flow of immigrants, in fact mass immigration was restricted to the 1948–51 period, during which the Jewish population doubled. Thereafter, the volume of immigration has fluctuated, never attaining the previous levels, but sufficient to make Israel a country of relatively rapid population growth, about 3–4 per cent annually after 1951. A commitment to unrestricted Jewish immigration involved a commitment to rapid development, recognized as the valid responsibility of government. Thus, most of the population has not looked adversely at the high level of government intervention, manifest in continuous exchange control, price control and rationing for several years, a relatively large budget, and direct public activities in many economic branches. But the way the government went about the mobilization and allocation of resources for absorption of the immigrants aroused continuous discussion.

An ambitious development program was undertaken. It was initiated by inflationary finance, but soon the major sources of financing capital formation were foreign aid, private and public grants, loans and investments. The resulting economic growth has been substantial: real GNP increased at an average annual rate of over 11 per cent and GNP per capita at over 6 per cent, despite two recessions in 1952–53 and 1966–67. Whereas unemployment was high in the early years, it started to decline in 1954, reaching virtual full-employment in 1965; after an increase in unemployment in 1966, unemployment is once again rapidly disappearing. In terms of growth of product, rising standards of living and attainment of full-employment, Israel's achievements have been truly remarkable.

Two shadows have been cast on these achievements. One is the almost continuous inflationary experience; the other is the continued current account balance of payments deficit, which though financed by substantial flows of unilateral and capital transfers, has been regarded by many as a problem. The persistence of these two problems, despite unceasing official commitment to mitigate if not eradicate them, has engendered continuous debate on their nature and the efficacy of economic policy.

* I am grateful to A. L. Gaathon, Harry G. Johnson and a number of my colleagues at the Hebrew University for helpful comments on an earlier draft of this survey.

[1] There are a number of general surveys of Israel's economic development in English, among them [120, 65, 162].

The policy issues surveyed in this article have been selected more for their importance in Israel's economic literature than for their intrinsic importance. The subjects considered are: economic planning, capital inflow and the import surplus, foreign exchange rates, inflation, and very briefly, agricultural policy.

Before commencing the actual survey, some comments may be made on economists in Israel. In the pre–State period Israel was not overly endowed with people concerned with economic theory and research, though some important research was done and a number of influential economists and many more successful men of economic affairs were at work, not a few of whom pursued their work after 1948. The economists among them were predominantly of Central European origin and training. With the development of the Hebrew University's Economics Department, a whole new generation of economists has emerged. Some of these have pursued advanced training abroad, but even the majority who are exclusively locally trained received an economics education which is basically what is standard economic thinking in the English speaking countries. The development of the Department of Agricultural Economics and economics departments at newer universities has continued and expanded this trend. The economic literature thus reflects the opinions of many active men of affairs, most of whom are self–taught or were formally trained many years ago, and those of many younger economists who, though differing in interests and opinions, share very similar educational backgrounds.

The economists holding positions at the universities have generally done post–graduate training in the United States or England; their interests are generally the same as those of their fellow students abroad. Some have contributed to the theoretical literature, and some publish abroad exclusively. Others have devoted much of their research to the Israel economy, each in the field of his theoretical interest. Since the first group is not represented in this essay, what must emerge is a biased view of economic thought in Israel. It is probably fair to say that the importance of immigration for Israel is reflected in the relatively large number of young economists interested in research on human resources. In research on income distribution, labor force participation, employment, consumption and saving habits interest is focused on differences between veterans and new immigrants and between immigrants from different continents. Much of this work is also unfairly ignored by this survey.

Economic debate in Israel takes many forms: all the daily newspapers give wide coverage to economic matters, there are several economic journals and annuals and there are frequent symposia and public discussions. This survey has entirely ignored the daily press, and concentrated on publications in journals and monographs; inter–departmental government memoranda are referred to only in selected cases, when the subject received wide circulation and aroused comment. One final introductory remark: the local literature is surprisingly extensive; references have been selected on the basis of historic priority or because they present good examples of certain opinions. No attempt is made to provide exhaustive coverage.

I. *Economic Planning*

Planning has been chosen as the first topic because it was the first major issue aroused by the need to cope with mass immigration, and because changes in the objectives of planning closely reflect changes in economic problems. The term economic planning means different things to different people. To some it means prep-

aration and implementation of detailed plans. Although agricultural settlement, water development and public housing—sectors very much subject to the control of government and public institutions—have been extensively "planned" in this sense, there has not been any such national economic planning. There has been considerable planning in the sense of forecasting desirable economic developments and pointing out the interrelationships between developments in various sectors and the role of economic policy. Actual developments often did not conform to these plans, for any or all of these reasons: the basic assumption of the plans proved unrealistic, the plans were poorly prepared, or insufficient efforts were made by policy makers to implement the plans. Finally, there has been considerable "planning" in the sense of adopting criteria for choosing between alternative development projects. This section considers separately long-term national planning, short-term national planning, and investment criteria.

A. *Long-term Planning*

During the pre-State period, it was felt that much faster economic development could be achieved if the Jewish Community had the tools to implement long-term plans. The Jewish Agency had a plan drawn up by Gruenbaum in 1946,[2] but could offer it only as a justification of its argument that Palestine could support large-scale immigration. When Israel gained independence, it was assumed almost automatically that development would be planned. The Prime Minister selected Gruenbaum (Gaathon)—clearly the most qualified man for the job[3]—to prepare a development plan. A four-year plan for 1950–53 was prepared [62]. The strong point of the plan was that it pointed up the interrelationship of the eight production branches, and final use sectors, and the input sectors. The main deficiencies of the plan—some of which Gaathon himself later discussed [49]—were in the assumptions, the targets, the available data, and most important, in its implementation.

Gaathon was told to build a plan on the assumption that 750,000 immigrants would be added to the population in the four years covered. In fact, net immigration was less than half that amount, and practically entirely in the first two years, thus, a crucial assumption proved correct for only the first half of the period. Over-optimistic estimates of immigration have plagued every subsequent plan as well, but the quantitative importance of the error was of course greatest in this first plan.

Gaathon was given several targets—goals whose attainment was quite unrealistic. Thus, he had to allow for full-employment, a curbing of the balance of payments deficit, and most unrealistic, the absorption of twenty per cent of the population in agriculture, all at the same time. He had no up-to-date data on national income, and none at all on capital and productivity. After processing the plans of all ministries through his intuitive grasp of the basic economic workings of the economy, Gaathon was able to come up with a plan which was not a bad forecast for the first two years [46].

The main failure of the plan was that no attempt was made to implement it; government ministries were unwilling to give up their prerogatives. The implications of separation between planners and plan implementers were understood by future planners; but they too failed to bridge this gap.

The desire to plan remained strong, and the need for planning became more acute

[2] L. Gruenbaum, *Outline of a Development Plan for Jewish Palestine*, Jewish Agency, Jersualem, 1946.

[3] Gruenbaum had not only already prepared a development plan, but in doing so had used the concept of a dozen sector input-output model which he had independently developed in the 1930's. (L. Gruenbaum, *National Income and Outlay in Palestine 1936*, Jewish Agency, Jerusalem, 1941.)

in the early 1950's, as foreign aid provided large sums for investment and the question of how utilize them became paramount. This was one of the main reasons for setting up an Economic Advisory Staff in 1953. This was a group of foreign and local economists, headed by Oscar Gass and Bernard Bell. However, Gass and his associates felt that neither the existing data nor the economic and institutional structure at that time were conducive to effective planning. Thus, though the Economic Advisory Staff did much to influence economic thinking, it did not prepare a new plan. Yet while with the Staff, Gaathon prepared a "model"—which he was careful to describe as neither a plan nor a forecast, but only a model to show the implications of a drastic reduction in capital imports between 1955 and 1963 [44]. Here too is evident an unrealistic target: the model considered the effect of a reduction of long-term unilateral transfers and net capital inflow to $50 million in 1963; the actual 1963 figure turned out to be $185 million. Considerable underestimation of future capital inflows have marked all subsequent plans as well. Gaathon also used his model to examine what would be the effects of a sudden increase of immigration to 100,000 in 1957 [45]. Considering reasonable changes in import components, and taking foreign aid as exogenous, he estimated the required increases in exports, and concluded—as did every subsequent forecast and model—that strong fiscal policy measures would be necessary to transfer the required amount of resources to investment and exports.[4]

In the late 1950's, interest in planning was once again aroused, this time mainly as a result of a belief in development planning evidenced by U.S. foreign aid officials. The Bank of Israel Research Department, with some help from the Ministry of Finance, started to prepare a five-year development plan. For this purpose, H. B. Chenery spent some time in Israel as an adviser. The results of this work were the preparation of a detailed input-output model of the economy, the preparation of annual forecasts-plans (discussed in the next section) and a long-term plan.

At this time there appeared a series of articles discussing the rational of planning, amongst which Guelfat [63], Merhav [103] and Kochav [87] deserve mention. Guelfat, starting from a Marx-Schumpeter development theory, assumed planning to be a necessity, and criticized the Bank of Israel's over-concentration on problems of inflation and money. He stressed that planning is more than a compilation of partial plans, and the consequent need for the public sector to take the initiative; the influence of the public sector would spill over into the private sector. Merhav also stressed the necessity of strong government initiative, but his main reason was that the undirected private sector could not be expected to invest in the industries with export potential. Kochav pointed out the distinction between long-term and short-term planning, the latter being limited by the existing economic structure, and the necessity of quantification of alternatives for both. Like Merhav, Kochav found the justification for national planning in the divergence between private and social profitability and the inability to rely on the distorted existing price system. He added the fact that the structural changes needed in the economy were major and not marginal. Accepting government intervention in the economy as necessary, Kochav argued that it would be pointless unless integrated by means of an overall plan.

[4] Thereafter Gaathon concentrated on improving the data on capital and productivity [48, 51], the absence of which had so hampered his earlier work [62].

The Bank of Israel's approach to planning was set forth in a series of working papers, and in two articles, by Zussman [151] and by Chenery and Bruno [30]. Zussman pointed out that the deficiencies of earlier partial plans lay not in their construction but in the absence of integration, and the main failure of Gruenbaum's plan was its divorce from the everyday economic activities of government. The new approach was designed to avoid these pitfalls. Government departments presented their plans and requirements; the planners examined these to see if they matched each other, in terms of input–output analysis, and if they added up to the goals set for the planners and the funds available for investment. A protracted give and take was envisaged, whereby the plans would be accepted as an integral part of each ministry's economic activities. Zussman also set out the main goal of the plan: to meet a decline in capital inflow to $200 million in 1964.[5] Assumptions as to population increase, changes in productivity and investment requirements led to estimates of GNP, required imports and necessary exports. Consumption was obtained as a residual.

A more formal development model was presented by Chenery and Bruno. Though their article presented only the aggregate model, the data were based on the multi-sector disaggregated input–output analysis. Their model considered three constraints on growth: the availability of labor, capital and foreign exchange. A major contribution of this article was to show how capital inflow affected two constraints: savings and foreign exchange. Accepting the pessimistic outlook for capital inflows (with some allowance for variation), they took into consideration several alternative assumptions as to increases in output per worker, the rate of saving, and the effective exchange rate.[6] The main point of the model was to show the policy alternatives that could be feasible, that is, would satisfy the equations of the model and fall within the predetermined limits as set by the various constraints. Both the Zussman and Chenery–Bruno articles stress the educational aspects of planning: showing policy makers the implications of various choices and goals and policies. There was renewed optimism about the political feasibility of planning, because unlike earlier plans and models, attainment of the set goals was now shown to be possible without curtailment of per capita consumption, but merely with moderation of the rate of its increase.

In 1962, the preparation of a national plan was transferred to a newly created Economic Planning Authority, set up first as part of the Ministry of Finance, and later as part of the Prime Minister's Office. The work of the Economic Planning Authority was in fact a continuation of the process described by Zussman: setting goals, receiving individual plans, checking their interrelationships and consistency with the aggregate model, and continuous feedback. The goals given to the planning authority were again the achievement of full employment with substantial growth, subject to the restraint of a drastically reduced capital inflow [161].[7] The actual preparation of plans during these years was not basically different from Gaathon's models; what was different was the level of sophistication made possible by a greatly broadened statistical base. In contrast to the early years, there were now available detailed estimates of national income and

[5] The tremendous increase in capital inflows in the 1960's was not foreseen at this time.

[6] Its effect was limited to increasing import–substitution and exports.

[7] Again, the expected reduction of capital inflow and annual rate of immigration were greatly exaggerated.

expenditure, a detailed input–output table,[8] Gaathon's estimates of capital stock, depreciation and productivity, and Liviatan's studies on saving and consumption patterns [163, 98]. The level of sophistication now made possible is illustrated by Merhav's study of how immigration affects the economy [104]. Merhav estimated the investment requirements for units of 10,000 immigrants, their addition to the labor force and effect on productivity, their demand for particular goods, the derived demand for investment in various branches, and the final effects on output and the balance of payments.

However, actual planning did not follow the direction set out by Chenery and Bruno. As mentioned, their article was concerned with pointing out policy alternatives; it concentrated on aggregates and used the interindustry analysis to make point forecasts. Bruno extended this work in a series of papers [25, 26]. He built a thirty sector programming model for the economy which focused on optimization of consumption.[9] The model solves for all shadow prices, including that of foreign exchange; the marginal productivity of capital inflow is measured in terms of increased output (or consumption). These models were static, focusing on a future point in time. Bruno, with Fraenkel and Dougherty [27], converted them to dynamic models, taking into account the time–path of development.[10] But the official planning has moved in the other direction: fixing major goals and grinding out their implications, without allowing for flexibility or even examining alternative assumptions as to goals or means.

The plan prepared in 1962 was not put into operation; a new Minister of Finance had assumed office, one less committed to the concept of overall national planning. In 1967, the work of the Economic Planning Authority was revived, and a new long–term development plan was prepared [162]. Though the planners followed the general method used for the earlier plan, they were less concerned with checking the internal consistency of the various plans by use of interindustry analysis, and less inclined to challenge the initial plans prepared by the ministries. Individual plans were correlated, within the overall restraints set out as policy objectives; a tremendous marginal saving rate was obtained as a residual. No specific policy measures to obtain such unrealistic measures were examined or even suggested. Like all the other plans (including Bruno's model), the latest plan is in real terms; the price level is not dealt with and no allowance is made for the possible effects of inflation or changes in relative prices, including the exchange rate.[11] Since like its predecessors this plan goes beyond the confines of planning government expenditure without wrestling with the basic question of how government is to direct the economy to carry out the details of the plan, there is little reason to expect a larger measure of implementation than was previously obtained.

B. *Short–term Planning*

Short–term planning is of course involved in all governmental budgetary activities, be it the preparation and implementation of the ordinary budget, the development budget or a foreign exchange

[8] For 1958; only in 1968 did the Bank of Israel start work on preparing a new table, for 1965/66.

[9] There is thus a switch in objectives, from pointing out policy alternatives to showing optimization at a future point in time.

[10] A basic problem of the official plans is that they assume steady development; for example, the average rate of GNP increase expected between 1963 and 1967 (about 10 per cent) turned out to be realistic, but this period consisted of two years of faster growth followed by stagnation.

[11] Bruno's models do allow for variation in the effective rate.

budget.[12] Improvements in methods of preparation and carrying out these budgets have been almost continuous. However, national planning for the coming year was not introduced until 1959, when a National Budget was prepared and presented to the Knesset (Israel's legislative body) along with and as background material to the proposed 1959/1960 Budget.[13] Subsequently, National Budgets have been drawn up every year, except for 1968.

The idea of a national budget is a sound one: given the tremendous relative importance of the government's budgetary activities, it is reasonable to present to the Knesset a document summarizing the economic developments of the past year and forecasting the developments of the coming year, stressing the role of the Budget and of government policy. It was also assumed that the annual planning would be part of the long-term planning, thus the National Budget could be used to show what progress was being made towards long-term objectives. This was emphasized in the first National Budget [157], and by Kochav [87].

The National Budgets estimate immigration and capital inflow, and on the basis of government policy and assumptions as to changes (or lack of change) in private expenditure patterns, arrive at forecasts of GNP, employment, the use of resources and price movements. Government fiscal and monetary policy are spelled out. As forecasts, these National Budgets have not been very accurate: immigration has generally been overestimated, the increase in imports, consumption and prices understated. But the substantial annual increases in GNP between 1959 and 1965 were enough to support the annual claim that "general developments" were as predicted, and gloss over the fact that very important aspects of the forecasts were wrong and in fact even policy variables were not correctly foreseen. The inaccuracy of this forecast became unpleasantly evident in 1966. The National Budget had predicted that the government's "restraint" policy, adopted in 1964 and slowly implemented in 1965, would lead in 1966 to a more moderate increase in GNP (7–8 per cent). When actual events showed that GNP was practically stagnant in 1966 (and per capita GNP declined), unemployment emerged as a major problem, and gross investment declined by 18 per cent instead of increasing 15 per cent, it was clear that short-term planning was deficient.[14] Yet this did not prevent the preparation of a National Budget for 1967. This too was a poor forecast—though the events of 1967 were sufficiently unexpected to excuse faulty forecasting.[15]

A model useful for short-term planning has been prepared by Evans [42] who spent a year in Israel to apply to Israel—with proper allowance for particular Israeli features—the Evans-Klein econometric model developed for the United States.[16] Unlike the Bruno models designed for long-term planning, this model relies mainly on regression equations estimated from time-series, and not on input-output analysis. The equations include price variables. Moreover, the model concentrates on developing multipliers to analyze the effects of changes in govern-

[12] The latter was introduced in 1952, on the suggestion of Oscar Gass, and used annually as a framework for authorization of foreign exchange expenditure until 1965.

[13] In 1958 Odd Aukrust spent some time at the Bank of Israel as a U. N. adviser, and helped in the operation of the National Budget.

[14] The arguments against the official view that the recession policy of 1965–67 was a grand success are presented in an article by Beham and Kleiman [14].

[15] However, the aftermath of the Six Day War was to increase economic activity; otherwise the forecasts would have looked even worse.

[16] M. K. Evans and L. R. Klein, *The Wharton Econometric Forecasting Model*, University of Pennsylvania, Philadelphia, 1967.

ment spending (of various types), changes in taxation and changes in exogenous variables such as immigration and foreign transfer payments to individuals. Using his model (based on 1952–65 data) to "predict" the *ex post* 1966 recession, Evans obtained better results than a naive model or the National Budget, but he of course included all the *ex post* exogenous figures. The Bank of Israel is currently working on the application of the model to short-term forecasting.

C. *Investment Criteria*

The problem of evaluating alternative investment projects, at a time when the price system was clearly distorted, was recognized in the early 1950's. Gaathon, Kessler and others were groping for suitable criteria. Kessler is credited with being the first to suggest viewing the official foreign exchange rate as the most "unrealistic" price, and trying to rank projects according to the real cost of net foreign exchange earned (or saved).

Barkai [5] was working independently on the same problem in 1954. He suggested several alternative methods for evaluating an economic project: (1) using the existing price structure to determine the comparative advantage of particular projects; (2) finding the project which produces foreign exchange at lowest cost in terms of the factor of production assumed to be scarce; and (3) estimating the cost in local currency (at market prices) of net foreign exchange earned (or saved). The three criteria he suggested were dictated by the fact that he considered various factors as incorrectly priced by the market; he did not suggest incorporating the various criteria in one measure, nor establishing some correction factor for converting market prices to real prices. His cost-of-net-dollars-saved criterion, while very similar to that evolved at the Economic Advisory Staff by Kessler and others, differed in one fundamental way: the usual method is to define the cost of foreign exchange as net local currency costs divided by net foreign exchange receipts (or savings, in the case of import substitutes); Barkai converts all foreign prices into local currency at the official rate of exchange, thus achieving the same result *only* if foreign exchange inputs are charged at the official rate. The problem of multiple rates for inputs was not a major concern at that time; Barkai was more concerned with the overpricing of labor at a time of high unemployment.

In 1954, Cats and Bawly started a research project to determine the value-added per man-hour in terms of international prices. Only partial results were achieved [28]. Cats again suggested this approach in 1961 [29]. Essentially, his idea is that capital commands a similar price everywhere, so that labor costs are the major international variant. He estimates net value added in foreign exchange (f.o.b. for exports and c.i.f. for import substitutes), subtracts an imputation for returns to capital, and divides the result by man-hours to produce the product examined.

In an unpublished master's essay Toren [140] discussed the necessity of finding a clear-cut criterion for evaluating investment projects—a necessity arising from the major role of the government's Development Budget in financing investment.[17] Toren pointed out the need to correct *all* market prices and to give quantitative values to frequently used intangibles such as defense considerations. Although he did not attempt to formulate rules for estimating these corrections or accounting prices, unlike Barkai he did show how they could all be brought into one criterion of profitability. He pointed out that if all other prices used were the same, the cost per dollar saved criterion

[17] The role of government in financing investment is discussed in [120] and [65].

would rank projects in the same way as his profitability criterion, but that his method alone could be used to evaluate projects that did not produce or save foreign exchange.

During 1956–59, several alternative methods were being used by the Ministry of Finance (generally by different departments). One was to use the subsidy rate for exports as a shadow price for converting foreign exchange to local value, and solving for a rate of return. Another was to use a shadow price for the rate of return on capital, and solve for the cost per dollar saved. A third, used especially for large agricultural investment, was to use the first method after allowance for various external benefits. Between 1959 and 1961, the Budget Division circulated a series of guidelines for evaluating projects: initiated by Toren, most were written by Shlein, e.g. [169]. He advocated the use of discounting to determine present value, or the internal rate of return. Projects ranked by the internal rate of return could be compared to the cut–off real rate of return, defined as the return on the marginal project for which budget funds were available. (The Budget Division guessed this figure as 8 per cent.) Shlein was careful to point out that this criterion alone could not be used to choose between mutually exclusive projects, and that total profits rather than rate of profit would have to be estimated in such cases. He also agreed that the same ranking would be achieved by use of the cost per dollar saved method, if the same shadow prices were used.[18] He preferred the rate of return approach because he felt this was conceptually more relevant to the question of pricing the scarce factor capital.

In a general discussion of criteria for selecting export and import–substituting projects, Bruno [24] shows how a programing model solves the problem of shadow pricing,[19] and how the cost per dollar saved criterion can be used instead, whenever lack of data prohibits more sophisticated studies, and where the rate of exchange is considered the price which should be solved as a residual. His 1958 input–output analysis for Israel presented the most detailed estimates of cost per dollar earned by exports, taking into account indirect inputs and price corrections for certain underpriced inputs [22, pp. 104–114].

II. *Capital Inflow and the Import Surplus*

Israel has been fortunate enough to be the recipient of very substantial annual inflows of unilateral transfers and long–term capital, which together came to a net figure of some $7,500 million for the 1949–67 period as a whole.[20] This inflow has financed a persistent deficit in the current account of the balance of payments, usually referred to as the import surplus. Virtually everyone who has written about the subject recognizes the tremendous importance of the import surplus for the growth of the Israel economy. However, the amount of research on the subject is less than its importance warrants. More attention has been devoted to the negative implications of the import surplus, what has been called the problem of "economic dependence."

Before proceeding, it is worth pointing out that much of the discussion uses the terms "capital inflow" and "import surplus" interchangeably, thereby introducing some confusion. In the purely techni-

[18] Here too, total gain rather than cost per dollar is relevant for mutually exclusive projects, as suggested also by Gilboa [54].

[19] His method takes into account the inter-relationship of changes, and does not examine a single investment project on the assumption that nothing else changes.

[20] This and other figures in this section are from [65, Hebrew edition, Ch. 8].

cal sense, the two are equal,[21] when by "capital inflow" is meant the net total of all unilateral and capital transfers, including short-term capital. But most Israeli economists refer to the long-term inflow only; thus, the "capital inflow" equals the import surplus only when net short-term assets neither fall (as they did each year in 1949–53) nor rise (as they have done each year in 1958–67). The distinction is important in another respect: the import surplus measures a real addition to available resources, whereas capital inflow refers to the financing of the import surplus and is often used to focus attention on the importance of the composition, sources and destination of this financing. The misuse of terms is not important when the context makes clear what is meant; but it may lead to misleading conclusions, particularly in discussions of economic dependence, when forecasts of the size of capital inflows are accepted as effective restraints on the future import surplus, and the availability of foreign exchange reserves is ignored.

A. *Effects on the Economy*

The effects on the economy of the import surplus or the capital inflow have been discussed by many. The most succinct analysis is by Michaely [106, pp. 40–51], who stresses the effects on growth, on consumption and saving, and on the structure of employment. This order is followed below.

Almost everyone agrees that the import surplus, which by definition increased total available resources, provided the resources for investment (since domestic savings were negative or zero most years), and that capital formation so financed was a major reason for the rapid growth of income per capita.[22] Patinkin [120, Ch. 3] and Michaely stressed the role of the capital inflow, particularly the fact that much of it went to the public sector (the government and national institutions) who used it to finance ambitious development budgets. Patinkin maintained that its absence would clearly have meant a reduced investment program.

"What might have been if" is a very perplexing question, which few have attempted to analyze. In two detailed studies of particular sources of capital inflow, one of German Reparations [155] and one of U. S. agricultural surpluses [55] Ginor has wrestled with this problem. In general, she took a maximalist approach, assuming (while recognizing this as an overstatement) that without reparations (and agricultural surpluses) imports and investment would have decreased by the full amount of such aid. She then estimated the effects on income, employment, income distribution, taxation and even inflation of the particular source of funds. In specific cases she examined the "maximalist" assumption and found it wanting; for example she concluded that, because in 1954 unilateral transfers were used to build up reserves and reduce short-term debt, it was unrealistic to assume that the absence of reparations would have meant an equivalent decrease in imports. But she made no quantitative assessments of the impact of reparations or the agricultural surpluses for less than "maximum" assumptions.

Another study of a particular type of capital inflow was made by Klibanski, for private foreign investments [85]. This subject, so often referred to, has been little

[21] Ignoring "errors and omissions."

[22] E.g. Patinkin [120], Michaely [106], Halevi and Klinov-Malul [65]. The later works rely on Gaathon's estimates of capital and its contribution to growth [48, 51]. It is worth pointing out that in contrast to productivity studies for the U. S., which show that productivity is the major factor explaining the increase in output per worker, the studies on Israel show that though productivity grew at about the same rate as in the U. S., increased capital is the main factor explaining growth.

studied. Citron and Kessler [31] examined investments made through the Investment Center in 1948–54, but concluded that the data were inadequate for any estimate and evaluation of the impact of foreign investments. Klibanski, using data obtained for the balance of payments estimates, was able to estimate private foreign investment and its branch destination for the years 1958–60, and the financial structure of firms receiving foreign investment. He did not attempt to analyze further the impact of this capital inflow.

Both Patinkin and Michaely have pointed out that the existence of net dissaving in the economy means that part of the import surplus went to increase consumption. Ginor, in an early analysis [38], suggested that the over–valued exchange rate at which the capital inflow destined for investment was converted to local currency led to the subsidization of consumption. Her study of the agricultural surplus program, in addition to estimating the effect of the program on total consumption (as did the study of reparations), also examined in considerable detail its effect on consumption of particular agricultural commodities.

The effects on saving are of two kinds. One is definitional. Creamer [33, p. 21] has shown that personal income was larger than national income in 1954, due to transfers from abroad; thus the definition of personal saving depends on the way foreign transfer payments are treated in the national accounts. Patinkin [120, p. 97] discussed the effects of alternative treatments of transfers both on the saving totals and on the allocation of savings between households and the public sector, when transfers to individuals are made via institutions. The dissaving sector is determined by the institutional arrangements for transfer of unilateral funds from abroad. In particular, he stresses the difficulties involved in deciding the extent to which recipients of transfers treat them as income in their decisions to save or consume. Another definitional problem arises from the standard practice of recording investment in human beings as consumption. It can be argued, and often has been, that aside from the general issue, a country receiving very large numbers of immigrants should consider much of the expense of their education and adaptation as investment.

More fundamental is the extent to which the availability of capital inflows reduced the level of domestic saving. Michaely suggested that the capital inflow, going as it did mainly to the public sector, made possible cheap loans to firms, reducing their need to provide their own capital, and thus business saving. Bank of Israel studies by Dovrat, Wahl and Tamari [37] on the financial structure of industrial companies support this contention. Halevi [65, p. 97] had made a more sweeping contention: domestic savings were zero or negative because the economy did not need them to finance investment. In other words, disagreeing with Patinkin's idea that had the government not received such a high level of capital transfers the entire investment program would have been substantially smaller, Halevi maintained that given Israel's commitment to rapid development spearheaded by government, a lower level of capital inflow would have necessitated a higher level of domestic saving. Of course the disagreement is really quantitative; investment would certainly have been less without an import surplus—savings sufficient to replace the entire import surplus would not have been generated—and the question at issue is how much less.

Michaely suggested that the import surplus may be a major explanation of a phenomenon often remarked upon: the high concentration of the labor force in services. His explanation is that since international trade is goods–intensive, rela-

tive to services, a country having a high level of imports, much of which is the result of an import surplus, would tend to concentrate production on services. This is a substitution effect. In addition, Michaely suggests an income effect, since the import surplus provides a higher level of total resources per capita—which may be a factor affecting the share of the labor force engaged in services.

Ofer undertook a study to determine the extent, if any, of over-concentration in services in Israel. Using international data to establish a correlation between income per capita and the share of the labor force in services, Ofer indeed found that Israel has an over-concentration in services [115]. He examined the influence of the import-surplus, and found that substitution effects and specialization effects of international trade explain more than half of this over-concentration. The income-effect provides an additional, though less important, explanation. However, the flow of foreign transfers to government explains, from the demand side, much of the over-concentration in government services.

B. *Economic Dependence*

As already mentioned, economists in Israel have been inclined to stress the negative implications of the persistent import surplus. The term "economic dependence" and ideas on how to achieve "economic independence" are part of almost every discussion of economic policy in Israel. The major contributions to the discussion have been by Creamer [33, pp. 22–23], Lerner [93], Gaathon [44], Patinkin [120, 121], Michaely [105, 106],[23] and Pines [123], and an analysis of the debate has been made by Halevi [65, pp. 167–171]. Though the discussants often fail to make a distinction, there are clearly two separate issues in the discussion of economic dependence; one may be called the "cessation problem" and the other the "future burden problem."

The availability of an imports surplus increases total resources, and its importance can be shown by the ratio of import surplus to total resources, i.e. to national product plus the import surplus. This index has been used as a main (if not always the only) measure of economic dependence by almost all the writers on the subject.[24] They see the contribution of the import surplus as a measure of what the economy must do without, or provide by other means, when the import surplus eventually declines as a result of a decrease of foreign transfers. This index is used by Creamer, Patinkin and Michaely to show the extent to which the economy has progressed or failed to progress towards economic independence. Whereas this index shows dependence, the extent to which changes in the index measure progress towards independence is highly questionable.

A question frequently but inconclusively debated has been whether the index should be computed in constant or current prices. Creamer used constant prices, expressed in dollars,[25] Michaely used prices expressed in local currency; and Patinkin used current prices. Patinkin and Michaely have debated the point.

Patinkin explicitly stated that changes in prices, expressing the relative increase in

[23] The issue was debated several times at the Hebrew University in late 1960 and during 1961. Michaely refined his views in a paper circulated in the Economics Department, but unfortunately not published. References to Michaely are based on this paper as well as the two cited.

[24] Patinkin inverts this index, showing GNP/total resources, thereby getting an index whose increase shows declining economic dependence.

[25] His method was to use an implicit price index to express national income in constant 1954 IL; to convert to dollars at the official 1954 rate of exchange—which he assumed to be "realistic"; and then to correct for changes in U. S. prices by using the implicit U. S. GNP price index.

the evaluation of imports, should be taken into account. Michaely and Pines, in independent papers, used similar assumptions to show that though neither index accurately measures changes in "real" economic dependence—when this is defined in terms of physical goods—the constant price index is a more accurate indicator of direction of change than the current price index. This is not necessarily true if the transformation curve changes other than by "natural growth," or if demand conditions change. Michaely admitted that a deterioration in the terms of trade should be taken into account, but he maintained that the main reason for the divergence between the two series, with the constant price index showing a larger decline in dependence than the current price series, arises not from changes in the terms of trade, but from changes in relative prices of imports and domestic goods in local currency, specifically from devaluation beyond rates suggested by purchasing power parity. These changes he believed should be excluded. This approach would be more acceptable if Michaely argued that such devaluation was excessive; but he does not: he has steadily maintained that devaluation must exceed that necessary to maintain purchasing power parity. It is thus not clear why he makes the distinction between changes in terms of trade and devaluation for the dependence index. A criticism of the Michaely–Pines approach implicit in Patinkin's formulation is that in a multiple good world, deflating by price indexes does not give a measure of physical goods, but only a measure in terms of some previous scale of evaluation; why should the past price ratios be considered more "real" than the present scale of evaluation?

The problem of cessation also involves a time dimension. If the import surplus were to decline abruptly, immobility of factors of production would cause a decline in national income, additional to the decline in the import surplus. Pines has shown diagrammatically how this decline is a function of the time allowed for adjustment to a decrease in the import surplus. Michaely has suggested using an index showing the ratio of the import surplus to total imports, or a variant, the ratio of exports to imports (a measure used by Ginor [41]) as indicative of the decrease in imports and thus of the economic dislocation which a sudden decrease in the import surplus would cause. But it is obvious that such an index, though useful, does not take into account the level of foreign reserves, which could be used to mitigate the effects on total imports of a sudden fall in the import surplus, and could allow for a more protracted investment program.

Most of the participants in the economic dependence debate make suggestions as to how to hasten the achievement of independence, though some, such as Lerner, visualized complete elimination of the import surplus, whereas others, such as Gaathon, Horowitz [75], Patinkin and various economic planners (discussed in section I above) would have been satisfied to achieve a smaller import surplus, equal to their estimation of "normal" future unilateral and capital transfers. Lerner, in a simplified diagrammatic presentation, emphasized the need to cut imports, consumption, and export prices; Patinkin [120, Ch. 5], in an arithmetic example, Gaathon [44] in a simple model and later planners in more sophisticated models have all believed that economic dependence could be achieved or approached without cutting either consumption or imports, but only their rate of increase.

These projections all assume a *gradual* decrease in the import surplus; such an implicit assumption underlies the use of *ex post* movements in the "index of dependence" for evaluating past economic policy. Patinkin [120, p. 132] maintained that

the failure to decrease economic dependence, as his index measures it, between 1953 and 1958 was the major failure of the Israel economy during its first decade.[26] The point that preparation for a future decline in the import surplus need not appear in an unique way in the standard index was made by Golomb [57] and expanded by Pines and Halevi: a given capital inflow can be used to increase investment increasing the current import surplus—or can be used to build up reserves, thus cutting current imports and lowering the index of dependence. In their disagreement with the use of indexes of past import surplus as measures of preparedness, they emphasized that in the 1960's the annual capital inflow was so large as to give considerable leeway in decisions to increase reserves or allow a more liberal import—and import surplus—program. Bruno, Fraenkel and Dougherty [27] have examined the time-path to economic independence, showing how a future output and balance of payments position can be attained by alternative patterns of growth.

What has bothered many Israeli economists is that the past and present high level of available resources is achieved at the cost of future debt and interest charged unmatched by productive investment. This idea underlies Bonné's contention [20] that much of the import surplus was "wasted" by not going to productive investment, and Abramowitz's claim [1] that the true measure of economic development is whether the import surplus went to investment or to consumption. Gaathon [44] expressed this more formally, by defining the existence of negative saving for the economy as a whole as a sign of economic dependence, and Patinkin and Michaely both used saving ratios as alternative indexes of economic dependence.

Halevi has pointed out that the high level of grants and gifts must be taken into account; thus for the future burden aspect both the import surplus/total resources index and the import surplus/investment index should include only that part of the import surplus financed by debt creating capital transfers, and exclude unilateral transfers.

Patinkin focused attention on the interest burden by also stressing the ratio of net interest payments abroad to the import surplus and to exports. Michaely added annual indexes of the ratio of net foreign indebtedness to the net national product, the ratio of foreign indebtedness to export receipts, and the ratio of foreign indebtedness to the national capital. He emphasized that each of his indexes illuminates a particular aspect of economic dependence and cannot be a single measure of it. Halevi suggested that these indexes are really attempts to substitute macroeconomic indicators for a cost-benefit analysis of foreign borrowing. Patinkin has stated [120, p. 128] that he would not be worried about the interest burden were he convinced that investment financed by marginal foreign loans were sufficiently productive to cover interest charges. He felt that no data were available to measure this, and the problem was further complicated by the fact that marginal activities financed by the import surplus went to increase consumption as well as investment.

This problem was first tackled by Bruno [25] in a planning model which attempts to measure the marginal productivity of foreign exchange in terms of increased consumption. In effect, Bruno's model is an attempt to measure the transformation curve showing the trade off between total resources and the capital inflow, what Pines had shown diagrammati-

[26] Michaely too stresses the importance of gradualism, but he points out that care must be taken in interpreting changes in the index; e.g. a reduction from 60 to 30 per cent is much easier to achieve than from 30 per cent to zero

cally. The Bruno, Fraenkel and Dougherty dynamic model [27] brings into the estimate of the transformation curve intertemporal choice between consumption and capital inflow levels. In this model the marginal productivity of foreign exchange at alternative levels of foreign borrowing is compared with an arbitrarily chosen interest rate of 6 per cent on foreign borrowing. Fraenkel has been attempting to extend the approach by allowing for a rising supply schedule of foreign borrowing.

III. *Foreign Exchange Rates*

Foreign exchange control has been an integral part of economic policy in Israel. The widespread use of quantitative restriction—gradually eased starting from 1956—combined with taxes and subsidies on foreign exchange transactions, and for several years even multiple formal rates of exchange, has given Israel a complicated multiple exchange rate system. There were several formal devaluations which gradually raised the price of one dollar from IL 0.33 to IL 3.50 twenty years later. Though formal devaluations were sporadic, the constant changing of taxes and subsidies and switching from administrative to fiscal controls made effective devaluation continuous.

There has been surprisingly little thoughtful writing and research on the desirability and economic effects of devaluation in Israel. There has been much oral discussion, and perhaps departmental memoranda, particularly in the periods before the formal devaluations of 1952–55 and 1962. Michaely has maintained, on the basis of his study of past devaluation [108], that Israel must constantly devalue because reducing the import surplus involves increasing real cost. This is part and parcel of his analysis of economic dependence, which he measures in real terms. Bruno, Fraenkel, and Dougherty arrive at a similar conclusion in their dynamic programming model [27]; they show that the real cost of replacing capital inflow rises. The policy implications of this are clear; but makers have tended not only to delay devaluation (at least formal devaluation), as long as possible, but also to hail every formal devaluation as the last. Most arguments about devaluation have usually concerned two issues: the desirability of a unitary rate as opposed to multiple rates, and whether it is best to achieve such a rate by formal devaluation or by other means. The latter issue concerned the ability of the government, given Israel's institutional framework, to prevent domestic prices from rising and so to make an effective devaluation merely nominal. Riemer [127] for example, maintains that the cost-of-living allowance system (discussed in section IV–D) is devaluation-defeating. Rubner [131, p. 215] used a Polak–Chang formula, showing the 1957 average rate times the U. S. consumer price index in 1957, as a per cent of the Israel consumer price index in 1957 (for both 1949 = 100) to conclude that devaluation between 1949 and 1967 was 58 per cent successful. Michaely [108, p. 96], using much more elaborate data, compared an index of the average effective exchange rate (of which more will be said below)—both to exporters and to importers—to the national product price index and to a purchasing power parity index, for the 1950–62 period. His various indexes show that the exchange rate increased 2 to 2.5 times as much as product prices, but almost all the increases were between 1950 and 1955. However, no study has yet been made of the macro-economic effects of devaluation.

The methods and objectives of exchange control in general have also received relatively slight attention. A public committee concluded in 1953 [170] that exchange control was still necessary, and there has been

no serious official rethinking on the subject since then. Rubner [131, Part B] presented a detailed description of and scathing attack on the entire system. The most recent survey of exchange control in Israel was made by Halevi [64].

Considerable attention has been given in the literature to two aspects of the multiple exchange rate system. The first is the effect of multiple rates on the recording of the national accounts, particularly on the evaluation of national product and saving, and thus on policy conclusions. The other is the effective protection of domestic production and promotion of exports. These aspects are considered separately.

A. *Multiple Rates and National Accounting*

Data on imports and exports compiled by the Central Bureau of Statistics are recorded in local currency values by converting foreign exchange values at the prevailing official exchange rate. The official rate was rarely the rate actually used for foreign exchange transactions: black or "grey" market rates, multiple formal rates, and taxed or subsidized official rates generally prevailed. It was quickly realized that the use of official rates for accounting practice could distort the national accounts, and limit their use as basic information for evaluating economic performance and formulation of policy. The basic problem is as follows: The gross national product at market prices (GNPm) is estimated as the total of private consumption (C), government consumption (G) and gross domestic capital formation (I) *plus* exports (X) *minus* imports (M). Since imports exceed exports, the way in which the import surplus is evaluated in local currency affects the estimate of GNP. Specifically, imports are part of the domestic use of resources (C, G, I) at prices actually paid for them by domestic users, therefore if the deduction of the imports surplus is based on computation at the official (underpriced) rate GNPm is exaggerated. Even worse, saving for the economy as a whole is estimated as the difference between GNPm and the total of C, G and depreciation (or by an identical procedure, as the difference between net capital formation and the import surplus). Using official exchange rates exaggerates saving. This is for Israel more than a technical point: the continuous effective devaluation accompanied by sporadic official devaluation meant that trends were distorted, and it was not uncommon to hear such statements as "saving is finally positive and increasing," at the same time as others were saying "saving is still negative and showing no improvement"—different policy recommendations would follow from what exchange rate was used to convert the import surplus.

The first attempt to correct for such distortion was made by Gaathon. In 1951, there was only one formal rate, but large amounts of imports were brought in under licenses granted "without allocation of foreign exchange," and actually paid for at much higher effective rates. In his estimates of national expenditure for 1951, [46] Gaathon reestimated such imports at a higher rate than the official one.

After the 1952 devaluation, several formal rates were in force, and taxes and subsidies were explicitly designed to effect devaluation. In 1953, Gaathon and Toren, then with the Economic Advisory Staff, were concerned with what taxes should be added to the formal rate to approximate the effective rate on imports.[27] In 1954 and 1955, Kessler (also with the Economic Advisory Staff) and Lubell (at the Falk Project for Economic Research in Israel) were working on estimates of national expenditures, and experimenting with alternative effective exchange rate con-

[27] They were stimulating the cost of imports in the CPI basket.

cepts [160, 100]. A minimum estimate was to take the average formal rate, and not the official rate only.[28] A variant used was to add the increase in customs duties and other import taxes after the devaluation. Another variant (previously used by Toren) was to add only taxes which went into extra-budgetary accounts, such as exchange equalization funds.

The idea of distinguishing between duties that are "normal" and those that are devaluation-substitutes persisted. Riemer [130] used it to estimate the effective exchange rate after the 1962 devaluation: his idea was to take the average tariff in other countries as a measure of "normality." Baruh (who had made estimates of affective rates for both Kessler and Lubell) spent several more years on such estimates. One of the variants in his estimates [8] also excluded "normal" duties: accepting the exchange rate prevailing on January 1, 1955 as representing purchasing power parity and the then prevailing tariff structure as "normal duties," he included in the effective rate only increases in import taxes after that date.

But this is only a variant; in his estimates of the effective rates he considers most applicable, and in his detailed measures of rate dispersion in 1955–61, Baruh departed from the idea of including only above-normal duties, and included *all* duties and other charges specific to imports (and all subsidies on exports). This view was ably expounded by Gaathon [47].

In their corrected national accounts estimates for 1950–62, the CBS[29] National Accounts Department (headed by E. Levy) switched to recording international transactions at effective rates [96]. Because the Bank of Israel continued to present accounts using official rates, considerable discussion was aroused. In addition to Levy, Baruh and Gaathon, Michaely [107] and Wahl [141], have contributed to the theoretical debate, and Michaely has produced the most exhaustive estimates of various alternative effective rates [108].[30] The views expressed are surveyed below in logical rather than chronological order.

Baruh and Gaathon advocate including all duties and charges on imports qua imports and subsidies on exports qua exports. Levy [96] included also purchase taxes on imports, regardless of whether similar taxes were applied to domestic goods, arguing that final users paid for imports market prices which include *all* taxes on imports. Michaely argued that the effective rate for imports should include not only all taxes but quota profits as well, since they form part of the price paid by final users of imports. This view is shared by Wahl. Michaely's estimates, however, do not include quota profits, for lack of data.[31]

As regards exports, Baruh, Levy, and Gaathon and Michaely all agree on using the effective rate of exchange for exports, which includes all subsidies on the value-added of exports[32] (exports minus import component), although actual estimates differ in coverage of subsidies and in computation of value added. Michaely in a footnote [108, p. 67] suggests that if price discrimination exists, i.e., if export goods are also sold in the domestic market at higher prices than are obtained for exports, the higher prices should be used for converting exports to local currency values.

[28] Only one of the formal rates set by the government for foreign exchange transactions was recognized as the official rate.

[29] Central Bureau of Statistics.

[30] Michaely also refined the estimates by separating from both imports and exports direct import components of exports.

[31] This not only underestimates the effective rates, but also distorts estimates of their changes; the switch from administrative to fiscal controls on imports (a process started in 1956) meant that the estimates for later years include taxes instead of quotas, i.e. a higher rate is estimated.

[32] This concept is discussed in greater detail below.

He does not compute such price discrimination, but Bruno did something similar, for 1958. In building an input–output table for 1958, Bruno [22] attempted to value exports at the same producers' prices as similar goods sold domestically. He therefore estimated the price differential, which he called a hidden subsidy on exports, and added it to the value of exports. Ostensibly, Michaely would support doing the same thing for computing exports at market prices, which would thus also have to include any purchase taxes on export goods sold domestically. Wahl has taken an even more extreme position: since she considers exports as producers of the foreign exchange used to pay for imports, she therefore suggests using the average effective exchange rate for imports as the proper rate of converting exports as well. The GNPm (and saving) measured by the various methods would rank in size as follows: official rate > Gaathon–Baruh > Walh > Michaely.[33]

Whether GNPm is estimated by the use of official rates or effective rates, with or without purchase taxes, gross national product at factor costs (GNPf) is the same—differences exist only as regards the stage at which subsidies are added and taxes subtracted to covert GNPm to GNPf, since some are considered part of imports and exports instead of taxes and subsidies. Of equal importance, this GNPf (or rather net national product) is conceptually the same as national income, estimated as a total of payments to factors. But adding quota profits to imports, as Michaely and Wahl suggest, would alter the conclusion since they form part of profits on the income side. In fact, both Michaely and Wahl suggest a computation of GNPf that would differ even more from the traditional method: they would use the average effective rate for exports to convert *imports.* Their reasoning is that exports finance imports, therefore the factor cost of imports is measured by the price of producing exports. (It will be noted that Michaely disagreed with Wahl on using similar reasoning for estimating GNPm arguing that for market prices, the prices of goods are relevant; for factor prices, the alternative use of factors.) This suggestion results from asking "what is the factor cost of imports?" Whereas this question is of great interest for trade theory and policy, it does not seem to be relevant for national income accounting. GNPf is a measure of what factors actually received, and should be equal to the total estimated on the income side. To achieve this, Michaely and Wahl would have to consider the differential tax on imports—imports multiplied by the difference between the effective rates on imports and exports—as a factor payment to government.

B. *Effective Protection*

The difference between official and effective exchange rates are the result of taxation or subsidization of foreign exchange transactions; they reflect a conscious desire to promote or hinder certain

[33] Levy's method would rank somewhere between Baruh and Michaely, on which side of Wahl would depend on the relative importance of quota profits, and the difference between the effective rate of imports and exports. The relationship between the different concepts may be clearer if expressed symbolically.
Let:

Rx = official rate of exchange for exports.
Rm = official rate of exchange for imports.
Sx = subsidies on exports.
Tm = tariffs and special levies on imports.
Pm = purchase taxes on imports.
Qm = quota profits on imports.
$A = C+I+G$

Then the following evaluations of GNPm are obtained:

Bank of Israel: $A+X\cdot Rx+M\cdot Rm$

Baruh–Gaathon: $A+X\cdot Rx+Sx-M\cdot Rm-Tm$

Levy: $A+X\cdot Rx+Sx-M\cdot Rm-Tm-Pm$

Michaely: $A+X\cdot Rx+Sx-M\cdot Rm-Tm-Pm-Qm$

Wahl: $A+(X-M)\dfrac{M\cdot Rm+Tm+Pm+Qm}{M}$

activities. Specifically, exports and import substitutes have been encouraged, the former by means of numerous and complicated subsidies, the latter by administrative and fiscal protection. In addition, both have been promoted by direct investments and loans from the government's Development Budget.

The subsidization of exports has been surveyed by Pines [124], relying on preliminary estimates and concepts discussed by Gottlieb [60] and Bahral [3]. In the early 1950's, subsidies were given to certain export items per dollar (or equivalent foreign exchange) gross receipts. In 1953, the PAMAZ foreign exchange retention scheme was introduced, whereby the value–added in foreign exchange could be used to import high value (otherwise restricted) raw materials for production to the local market. Pines continued Gottlieb's estimates of the subsidy given in such cases. Both he and Bahral showed that the system was in fact one of subsidizing value added only, and as such preferable to subsidization of gross export receipts.

In 1956, direct subsidization of value added was introduced. The Ministry of Finance and Commerce and Industry examined individual product lines to determine the value added in foreign exchange, which alone received a subsidy. The method of computation was essentially the one discussed above; i.e. divide all local currency costs by net receipts in foreign exchange, though the extent of allowance for indirect import components was limited and arbitrary. By 1960 most goods were receiving a subsidy of IL 0.85 per dollar of value added (in addition to the official rate of IL 1.80 per dollar), although there were exceptions, both favored and discriminated against. In 1957–60, the IMF representatives, in their annual consultations, objected to the multiple–currency practices involved in the system. Their objection was not primarily to the fact that some exports received more than a total of IL 2.65 per dollar for net foreign exchange receipts, but rather to the unitary subsidy on net value added, which they saw as multiple rates on gross dollar receipts. In vain did Israeli economists try to explain that a uniform subsidy on value added was the best alternative to a devaluation, as regards exports. Of course, the problems of computation and the administrative difficulties of the system prevented it from being ideal for creating a uniform rate, a point illustrated by Pines.

There has been some discussion in Israel of whether the cost of value added is important or whether the proportion of value added also matters. Many have suggested supporting exports that have some minimum value added percentage, e.g. Ginor [39] and Goren [59], in fact, this was a criterion used for several years. Rubner [131, p. 127] has ridiculed the idea. Although the per cent of value added should not be the basis for different subsidies per dollar of value added, what Ginor and Goren had in mind was something else: the danger of fluctuations in prices abroad can easily turn low value added exports into negative earners. Moreover, as already mentioned, in the selection of mutually exclusive projects the total contribution of foreign exchange is not irrelevant.

The protection accorded import–substitutes had, until fairly recently, received less attention than the subsidization of exports, despite the fact that policy has certainly promoted import substitutes more vigorously than exports—a point stressed by Michaely [106]. Administrative protection was given almost automatically; only in the mid–1950's were considerations of price and quality given consideration in protection policy. The use of tariffs in conjunction with quotas was examined by

Gafni, Halevi and Hanoch [52],[34] but this study was concerned with locating protective tariffs and made no attempt to estimate the height of effective protection. In 1955, the economists of the Customs Division were already applying the idea of effective protection in their examination of requests for protective tariffs, but it cannot be claimed that decisions were influenced by the estimates provided. Goldberger (Ophir), in a master's thesis prepared at the request of the Customs Division [56], presented a clear application of the foreign exchange value added principle to import substitutes, and showed how duties on import components should be taken into account in estimating the effective protection rate[35] of a customs duty on a final product.

Whereas the conceptual problems were formulated at a relatively early date, actual estimates of effective protection rates were delayed for a number of years. The statistical difficulties are greater than those involved in estimating the effective subsidy given to exports—though the two concepts are basically the same.[36]

Michaely's study of effective exchange rates [108] also contains estimates of the effective protection rates, for 1956–60.[37] These estimates, though the best yet made, are deficient in two respects: they do not include quota profits in the effective exchange rate, and for purposes of finding the average effective rate for commodity groups items are weighted by the volume of imports, instead of by the volume of production so protected.[38]

Part of the new economic policy announced in 1962 was a decision to switch from administrative to fiscal protection. A public committee was set up to examine individual industrial items and decide whether to stop administrative protection and if so what tariffs to levy instead. After the committee had been working a year, Barkai and Michaely [7] criticized it for setting the rates of protection too high. Mandelbaum and Naveh [102] in turn criticized Barkai–Michaely, maintaining that they had over–estimated the effective protective rates since they under–estimated the rate of duty on imported inputs. A detailed estimate of the work of the committee through 1965, the relative importance of items examined and items subjected to fiscal protection, and the tariffs imposed, was made by Davidov [34]. She did not estimate the effective protective rates. Such estimates are currently being made in studies by two Hebrew University graduate students: Yoran is estimating the actual effective protection rate currently in effect, both for final products and for particular branches, while Tov is examining the protection rate implicit in the public committee's decisions and the reasons for the great dispersion in rates. In other words, Yoran is studying the results of protection policy while Tov is examining the implicit intentions of policy. A recent decision to institute annual across–the–board tariff cuts has aroused new interest in the subject, which will probably remain a popular subject of debate for some time to come.

A more fundamental aspect of the protectionist and export–promotion debate has been the question of unitary versus multiple effective rates. Bahral [3], Goldberger (Ophir) [56] and Pines [124] were primarily concerned with showing the effects on the allocation of resources of multiple unrealistic exchange rates, a subject also discussed by many others, e.g. the Foreign Exchange Committee in 1953,

[34] A summary of the approach and main results is presented in [53].

[35] The effective protective rate is not the same as the effective exchange rate on imports.

[36] Particularly difficult is the estimation of the foreign exchange value of domestic production and the quota profits on administratively restricted imports.

[37] He assumed the import component data of the Bank of Israel's 1958 input–output study were appropriate for two years before and after.

[38] For getting average import component, the weights used are the values of domestic production.

[170], Riemer [130] and Rubner [131]. These proponents of a unitary rate of exchange—effective if not official—are not always either explicit or in agreement as to what the unitary rate should be. Generally, it has been argued that the unitary rate should be the "real" rate of exchange. Most users of this term have meant an equilibrium rate of exchange, some estimate of the rate which would be set by a free market. Most have also meant the equilibrium rate at the time discussed, though others, such as Bruno [22, p. 106] had in mind a future equilibrium rate as most relevant for investment decisions.

Bahral, in a systematic and lucid application of price theory to Israel's multiple rate system, introduced to Israel a new concept of the "real" rate of exchange. After careful discussion of the demand for and supply of foreign exchange value added, Bahral tried to make the analysis more relevant to actual policy by allowing government restrictions of private demand. Quotas and tariffs are the way private demand is subjected to correction by government. The demand curve for foreign exchange as influenced by government is thus to the left of the free market demand curve. A lower equilibrium price is attained—though this is not a free market equilibrium price, it does equate supply and demand. Bahral maintained that this "real" rate should be used for all transactions; i.e. subsidies and taxes should be used to make it the unitary effective rate, thus equating the marginal cost of foreign exchange in all exports and import substitutes.

There seems to be a basic contradiction in this concept of the "real" rate and its application. If quotas are retained, some import-substitutes receive a higher effective protection rate than others (or than exports), and there is a misallocation of resources. On the other hand, if all quotas and special taxes (above those needed to reach the real rate) are abolished, then the demand curves "expressing the wishes of the government" which were used to estimate the real rate of exchange are no longer relevant, and the new "real" rate, which should be used for all transactions, is actually a free equilibrium rate.

For many years, the Ministry of Commerce and Industry was severely criticized for maintaining multiple effective protection and subsidy rates. It failed to reply to these criticisms in the terminology of professional economists.[39] In their response to criticism by Barkai and Michaely, Mandelbaum and Naveh [102] rephrased the justification for discriminatory pricing. After restating the frequently voiced justification of protection against brand name competition and dumping, they present the standard infant industries of developing nations arguments, which are essentially assertions that existing comparative advantage differs from long-term comparative advantages, and that growth can be speeded up by discriminatory protection. Strangely enough, this argument has not been followed to its logical conclusion—that exports as well as import substitutes should receive special treatment; the argument has been used merely to justify gradualism in reducing protection to some uniform level. The market versus bureaucratic appraisal of social profitability controversy may be expected to recur and persist for years to come.

IV. *Inflation*

A. *The Inflationary Experience and Policy Issues*

Except for brief respites, Israel has been subject to inflationary pressure since 1948. This inflationary pressure has found expression in rapid price and wage increases, frequent devaluation, and a persistent current account balance of payments deficit. In the early years, inflation was clearly of

[39] Lower subsidies to certain exports—mainly citrus—had been justified on the ground that the supply curves are inelastic.

the demand-pull type, marked by tremendous government deficits and a rapid rate of increase in the money supply. Price control and rationing were employed in the 1949–51 period. In 1952, a New Economic Policy was implemented, consisting mainly of devaluation, a raising of the domestic price level to soak up accumulated inflationary pressure, and efforts to curb the government deficit and the increase of credit. This policy was carried out over a three year period, by the end of which the Bank of Israel had been set up, and responsibility for monetary policy transferred to it from the Ministry of Finance.

After the first period of suppressed inflation there have been several years of demand inflation, several of cost inflation, and a number during which inflation was a mixed variety where the causes are less easy to identify. Attention fluctuated between demand and cost factors—among the latter especially the cost-of-living allowance which is an important part of wages—and consequently monetary, fiscal, wage and price control policy[40] alternated in arousing discussion.

Inflation aroused more discussion than any other topic. The Israeli literature consists of some purely theoretical discussions or reflections, and more frequently of arguments bolstered by selected data. Much less frequent are the more empirical studies. Since the Bank of Israel was set up in 1954, annual and occasional surveys of inflation and its manifestations have been available; these have been the basis of much of the discussion. In addition, there have been a number of general surveys of the inflationary experience during particular periods, among them: Gross [61], the first study of inflation in Israel, restricted to the 1949–50 period of repressed inflation; Patinkin's two surveys [119, 120, Ch. 4], the first covering 1949–53 and the second 1949–58; Kochav [88], covering the 1955–60 period, and the most recent, Klinov-Malul [65, Ch. 11],[41] covering 1949–65 (1949–67 in the Hebrew edition).

The discussions surveyed are organized around the following subjects: the manifestation of inflation, demand inflation and the role of money, cost inflation and the role of wage policy, and finally linkage schemes to mitigate the undesirable effects of inflation.

B. *The Manifestation of Inflation*

Basic to any consideration of anti-inflation policy, is the necessity to identify the existence of inflation. Much of the discussion of this question is connected with arguments about the causes of inflation, and as such is deferred to the next sections. Here, some approaches to the identification problem are considered.

In the early years (1949–51) inflation was identified with rising prices. This is the conclusion of Gross [61] who examined policy declarations during the period. Weiss [142], surveying price control, describes the attitude that held that control of prices was "war against inflation," and lowering prices a victory in that war. In later years, such attitudes were less crudely expressed; but official pronouncements of satisfaction with price stability in particular periods, e.g. 1959 and 1967–68, were often phrased in terms suggesting that absence of price increases means absence of inflation. This view has been stated explicitly by Tenenbaum [159], who maintains that there was no inflation in Israel in 1959.

The objections to this approach have been of two kinds. One focuses on inflation-

[40] Overt price control was gradually eliminated, with very few exceptions, but indirect control of prices have often been a major part of anti-inflationary policy.

[41] This is a co-authored book, with Halevi and Klinov-Malul sharing responsibility; however, references to a specific chapter are made to the author primarily responsible for that chapter.

ary pressures as opposed to price rises. Gross [61] and Patinkin [120] criticized the 1949–51 policy for attacking symptoms instead of causes, and viewed the creation of excess demand (which they identified with monetary expansion) as inflationary, despite the fact that controls prevented prices from rising *at that time*. Thus Patinkin viewed the large price increases in 1952, the result of the New Economic Policy, as anti-inflationary, since by lowering real money balances they decreased demand. M. Kalecki, who visited Israel briefly in 1951 as an adviser to the Ministry of Finance, advocated price controls; but he included in his report [79] recommendations for fiscal policy to reduce inflationary pressures. The idea of lags between the creation of inflationary pressures and price increases even in periods of less overt price control was stressed by Barkai and Michaely [6]. The idea of lags is still often ignored; 1969 is being hailed as a year of stability, despite great inflationary pressure.

The second objection to the identification of inflation with price rises suggests another way in which inflationary demand can expend itself: imports. That increased imports can substitute for price increases has been pointed out by many, e.g. Horowitz [72 and 75]; some have considered the import surplus as a significant variable. Patinkin [120] maintained that even during a period when certain outward manifestations of inflation, such as rates of increase of money, credit and prices showed a decline, the persistence of the balance of payments deficit proved that policy was unwilling or unable to cope with the fundamental cause of inflation. Others have restricted themselves to the view that an increase in the import surplus, rather than a failure to decrease it, is a sign of inflationary pressure. On occasion the Bank of Israel has used the change in the import surplus as a criterion, e.g. [156]. Klinov-Malul uses the increase in domestic resources at current prices as a measure of change in effective demand, this change being absorbed by changes in real growth of GNP, in prices and in the import surplus. The last two are her indicators of inflation.

C. *Demand Inflation and the Role of Money*

The role of money and consequently the role and effectiveness of monetary policy have been widely debated. The central position of the Bank of Israel in monetary policy and debate is such as to make it worth while to start with the main views of the Bank, particularly of its Governor, Mr. Horowitz, and then survey the discussions of particular points.[42]

Although there are in print alternative variants of and even some departures from what may be called the "Bank of Israel approach," the main idea is a simple application of the quantity theory of money, along the following line of reasoning: An increase in the quantity of money creates and measures an increase in effective demand [76, 71]. In the short-run, output is determined by real factors, such as the existing amount of factors of production [156]. Therefore an increase in effective demand can increase output only to the limit set by the supply factors. From this is derived the conclusion that the supply of means of payment should be allowed to increase only in the same proportion as the expected increase in output, or the expected increase in output plus imports or net imports [71, 73]. If the money supply increases by more than this amount, it will lead to price increases or balance of payments deterioration or both. This is the rationale for curbing the supply of money. The Bank, however, realizes that the

[42] Arguments concerning particular tools of credit policy, e.g. criticisms of excessive reliance on reserve ratios and little use of open market operations, are not surveyed here.

money supply is affected not only by credit policy, but by factors less subject to the Bank's control, such as the public's readiness to hold monetary assets in less liquid deposits, and primarily, fiscal policy [72, 74].

The Bank has also presented, on occasion, a more sophisticated approach. If the possibility of cost inflation is recognized, an increase in money supply may be a result rather than a prime mover. But whether it is inflation-causing or inflation-induced, an increase in the money supply is a necessary condition for inflation [79]. Unfortunately, the Bank often cannot avoid submitting to institutional pressures, and must increase the money supply to prevent unemployment [73].

The Bank has therefore consistently attempted to curb the money supply—and has been singularly unsuccessful most of the time. The discussion of the role of money and monetary policy can be related to several issues raised by the Bank of Israel approach: (1) Is the crude quantity theory applicable; (2) how does an increase in money increase effective demand; and (3) does the restriction of increase in the supply of money curb the growth of output? Patinkin [120] and Kochav [88] have both found that after 1954[43] the income velocity of circulation[44] was constant. Patinkin [120] explains this constancy, in an approach credited to Phillip Cagan, by the fact that a steady rate of increase in prices during this period meant that the cost of holding money balances was unchanged. The constancy of the velocity of circulation is used to explain that monetary developments in Israel "have been closely in accordance with those predicted by the crude MV=PT of the quantity theory of money." [p. 116]. But Patinkin does not suggest that monetary expansion was always the cause of inflation. Moreover, even granting that V remained constant—a debatable point—the large variations in annual increases in total resources (T) during the period discussed suggest that the crude quantity theory could have been used as a reliable guide for *ex ante* monetary policy only if accurate projections of changes in T could have been made.

Many have denied that velocity has been constant in Israel. Some, such as Tenenbaum [159], Bar-Yosef [9, 12], Lehman [92] and Halperin [66, 68] have claimed that V is not constant, but is affected by the rise of near-moneys, such as post-dated checks and bank negotiated bills, created as substitutes for money in periods of tight credit control. They refer to demand deposit velocity, not the income velocity of money. Whereas Halperin is not sure whether near-moneys are complete substitutes for normal bank credit, Bar-Yosef concluded that credit control is a failure, since the market creates substitutes. (Ben-Shahar's empirical study [16] also concluded that control of the volume of short-term credit was ineffective.) The main ideas here are that: (1) exclusion of near-moneys makes the official estimates of the money supply meaningless, and (2) bill-brokerage transfers purchasing power from low-velocity holders to high-velocity holders. The Bank of Israel [156] maintained that "do it yourself money" should not be added to the money supply, but it admitted that it does affect velocity. The large volume of such moneys created during the 1960's affected velocity greatly, as shown by Steinberg's estimates [137]. Those citing the change in velocity as evidence against the tight-money policy failed to offer proof that changes in V were sufficient to compensate for changes in M. Dovrat, in a recent article [36] claimed that changes in V did compensate in 1968 for increases in the money supply. He then concludes that

[43] Patinkin refers to 1954-58, Kochav to 1954-60.

[44] Defined in terms of total resources rather than income.

the oft-maligned MV=PT equation is really very useful. He fails to discuss the question of how the monetary authorities can anticipate large variations in V, and thus use the formula for their decisions.

Halperin raised an intriguing point [68]. He questioned the standard idea that reserve-creating demand deposits are really more high-powered money than cash held by the public. True, current account deposits can lead to expansion of the money supply; however, he believes that the velocity of cash is many times greater than the velocity of demand deposits, so that a shift from cash to deposits has a net deflationary effect! Unfortunately, his estimates of the two velocities were not entirely convincing, as pointed out by Arian [2].

A second controversy over the applicability of the crude quantity theory concerns the relationship between changes in the volume of money and changes in the import surplus. As already mentioned, the Bank of Israel and others have often pointed out that inflation can appear as increased imports rather than as price increases. Horowitz [75] found "by inspection" of means of payment and import surplus data a "visible correlation" of the two, without regard to the sequence which connects them. Tenenbaum [159] denied seeing this correlation; he saw no necessary connection between imports and inflation. Ben-Shahar [15] suggested a statistical test. He ran some correlations and found a weak positive correlation for 1951–60, and a weak negative correlation for 1955–60, from which he assumed a strong positive correlation for 1951–54. Despite the weak statistical evidence, Ben-Shahar supported the Horowitz view—based on a simple quantity theory which he accepted. What he presents is a quantity theory formula where $MV=P(Y+D)$, which can be written $D= V\ M/P—Y$, where Y is real GNP and D is the real import surplus. If V and Y are constant, obviously D varies with the real money supply.[45] Ben-Shahar concludes that the data do not show this because V and Y may have changed. He does not defend his implicit assumption that both are independent of changes in M/P.

Empirical studies concerning the relationship between changes in money and in prices have been surprisingly few—other than those "empirical studies" which consist of presentation of two annual series and derivation of conclusions by inspection. Evans [42] ran several regressions, using annual data, and found little relationship between money and prices. However, since this subject was relatively unimportant for his purpose, he did not attempt to introduce lags or other variables that would show a clearer relationship. Kleiman and Ophir have used monthly data and introduced various lags. Their preliminary findings [84] show that there are discrete lags of six months in the 1955–59 period and eight months in 1960–65 between changes in the money supply and their effects. In the first period, changes in money explain 19–24 per cent of monthly changes in prices; changes in the import surplus and in money combined explain 60 per cent of the quarterly changes in prices. For the second period results are inconclusive.

Much of the above discussion has implied that increases in money increase demand but have not examined how this is brought about. The opinions expressed on the subject have varied from simple statements with no or with only implicit theoretical backing, to sophisticated formulations of monetary theory.[46] The Bank of Israel position, as already discussed above, was inconsistent, ranging from attributing

[45] The use of the same price level to deflate GNP and the import surplus ignores the possibility of changes in relative prices of domestic and imported goods.

[46] Monetary theory as such is not the subject of this survey, therefore contributions by Israelis, mainly Patinkin's well-known works, are not discussed except as they affect the discussion of Israel's experience.

an automatic causal role to increase in the money supply to the admission that changes in money may be responses to other factors. The Bank has complied a flow of funds data and discussed them, starting with the 1959 Report, in an effort to point out demand-generating sectors. In the earlier Reports, an extreme position is taken; for example: " . . . every pound of credit enables the receiving sector to expand its demand without a corresponding increase in supply" [153, p. 206]. In later reports, a more agnostic position is taken, e.g. "In many cases it is difficult to establish a causal connection between the creation and financing of a demand surplus" [154, p. 428]. Perhaps this change in attitude reflects the fact that in recent reports Heth has been responsible for the flow of funds analysis; thus an even more skeptical view is presented in [70] and in [162, p. 218], the latter being a survey of monetary developments prepared for the Economic Planning Authority by the Bank of Israel.

The flow of funds data point up a distinction frequently ignored in Israel; that between increases in the money supply which are an integral part of increased demand, and those that affect demand less directly. Thus Barkai and Michaely [6] and Klinov–Malul [65] have stressed the difference between the government's demand surplus and the liquidity surplus, the latter affecting demand via the interest rate and the quantity of real money balances.

The rate of interest has aroused much discussion in Israel, mainly because of a maximum interest law and of subsidization of loans to selected sectors by the government. Much of the debate, e.g. [164], and some empirical work, e.g. Ben-Shahar [16], relates to discriminatory rates and their effects on the allocation of investment, but the effect of interest rates on demand has not been neglected. In order to circumvent the maximum interest restrictions, it has been customary to make extra charges such as commissions; Bar-Yosef [10] stressed the necessity of adding all such charges to the nominal rate of interest, and Ben-Shahar [16] included these and linkage provisions in his estimates of the "effective rate of interest." Ben-Shahar also estimated "real rates of interest," which are effective rates deflated by the relevant price indexes, a procedure suggested by many, e.g. Bar-Yosef and earlier, Sharon [135].

The local literature is replete with articles showing that interest rates are "high" and therefore contribute to inflation by raising costs; the proposed solution is an increase in the supply of money. There is little evidence to suggest that the no less frequent rebuttals, such as those by Kerem [82], Patinkin [118], and Sheffer [136], pointing out that money differs from other goods in that increasing its supply can lead to increases in the demand for it rather than to a lowering of its price, have done much to change opinion. A case in point is the reluctance to abolish legal restrictions on the height of interest rates, despite the almost unanimous accord among local economists that the "interest law" is detrimental, e.g. [164].

Barkai and Michaely [6] believe that raising the interest rate would have reduced investments and slightly reduced private consumption, at least in the inflationary period following the 1962 devaluation. Kerem [82] believes that within the relevant range of interest rate variation in Israel demand for money and the volume of investment are inelastic with respect to interest; however, he thinks that savings are more responsive to changes in interest. Though the role of interest rates in the securities market and their affects on borrowing from provident funds have been examined by Sarnat [133, 134], no detailed empirical study of the influence of

interest rates on either savings or invest ment has been carried out.

As could be expected, real money balances play an important role in explanations of the influence of money on demand by Patinkin and his students. In his first analysis of inflation in Israel [119], Patinkin stressed the effect of changes in real balances on the demand for goods, both when real balances increased during the repressed inflation of 1949–51 and when they decreased following the New Economic Policy of 1952. Clearly, the tremendous price increases which decreased real balances in 1952 and 1953 absorbed inflationary pressure. But these balances were not equilibrium real balances, they were created as a result of rationing and price control; they were forced savings. Therefore, neither their creation nor reduction involve the "real balance effects" of monetary theory, the kind Patinkin refers to in his analysis of the velocity of money during 1954–58 [120].

Barkai and Michaely [6] also stress real balance effects in their discussion of monetary policy, or rather the failure of monetary policy, after the 1962 devaluation. This analysis is in part an application to a concerete case of devaluation at full employment[47] of the reconciliation of the "absorption" and "elasticity" approaches to devaluation that Michaely had discussed in a theoretical note.[48] Concerned as they were with the tremendous increase in money following the devaluation, resulting from unilateral transfers to individuals and from conversion of foreign exchange deposits created by such transfers in the past—analyzed later by Beham [13]—they felt that their analysis was sufficient to justify stringent monetary and fiscal policy.

Though there has been no research specificially designed to measure real balance effects—surprisingly, considering Patinkin's great influence on an entire generation of Israeli economists—Patinkin himself has applied to this question the research of others on the use of the very substantial foreign transfer payments to households received as restitution payments from West Germany. The way households used these funds has been examined, starting with the first Survey of Family Savings (1957/58).[49] Landsberger made a thorough statistical study of the use of restitution payments in a doctoral dissertation [91]; some of his results been published elsewhere [90]. What all these studies show is that lump-sum payments are treated differently by recipients from regular income receipts as regards consumption of durables, and the form in which savings are held. Patinkin [122] has used Landsberger's findings to illustrate the wealth effect of increases in money. Whereas his previous work referred to real balance effects on current consumption, Patinkin here admits that such effects are relatively slight, and assert that the main effects of an increase in real balances is on the way wealth is held and thus on the demand not only for monetary assets but also for fixed assets, investment and durable consumer goods. Patinkin's point is that during the lag between an increase in money and the resulting rise in prices, monetary wealth increases, and there en-

[47] "Full employment" does not exactly apply to a situation in which GNP increases at 10 per cent per year.

[48] M. Michaely, "Relative-Prices and Income-Absorption Approaches to Devaluation: A Partial Reconciliation," *American Economic Review*, March 1960.

[49] This survey was supervised by a committee headed by N. Liviatan; the results were published [163] as a joint report of the Falk Project, the Central Bureau of Statistics, the Bank of Israel, the Institute of Applied Social Research and the Economics Department of the Hebrew University. The Bank of Israel *Annual Reports* present the findings of this and later surveys in their chapters on savings. N. Liviatan has also used the saving survey data to test the permanent income hypothesis [97].

sues a dynamic process during which demand for durable goods increases [122, p. 13]. Granted that an increase in wealth affects the demand for durable goods in addition to the usual affect of an increase in permanent income on consumption, the crucial question is to what extent a particular increase in the supply of money, and thus in real balances, is to be considered as an increase in permanent income or as a temporary increase in wealth.

In a study of fluctuations in the prices of land in Israel, Borokhov [21] also attributed major importance to increases and decreases in real money balances. However, his analysis rests primarily on the diversion of purchasing power to fixed assets as a result of controls in the consumer goods market. Here again, as in Patinkin's analysis of the 1949–51 period, does one observe genuine "wealth effects" of real balances or the monetary manifestation of an increase in income which was not allowed to be spent normally?

Some of the differences in policy conclusions arrived at in the various discussions of money and velocity stem from looking at different kinds of money and the effects of changes in their quantities. Patinkin concentrates on outside money and assumes no redistribution effects of inside money. Bar–Yosef and Halperin are actually concentrating on the possible distribution effects of inside money. For policy purposes, as opposed to its presentation of detailed statistics on means of payment and deposits, the Bank of Israel ignores the distinction between inside and outside money; it looks only at the total money supply.

All the quantity theory approaches mentioned above assume that an increase in money is inflationary if it increases demand more than supply can increase. The arguments concerning the relationship between changes in money and changes in output are thus mainly the result of differences in evluations of the employment situation at a given time, specifically whether and to what extent an increase in effective demand can increase output.[60] Frequent arguments to the effect that a shortage of money curbs output and that "the needs of business" must be met, e.g. Zak [149] and Bar–Yosef [11], persist despite many rebuttals, e.g. Patinkin [119]. Schweitzer [138] presented a more reasoned analysis of the possibility that in 1954 a shortage of credit, rather than of labor, was a constraint limiting production; a point disagreed with by Kochav [86]. Their debate is more memorable for being one of the few in which the possible beneficial effects of inflation are considered—with Schweitzer suggesting than an annual price rise of up to 5 per cent would stimulate investment and thus the growth of the economy, and with Kochav dissenting vigorously.

The idea that there may be a trade–off between the rate of growth of income and the rate of increase of prices or the import–surplus, and that maximum growth is not necessarily optimum growth, has been frequently stated or implied, e.g. Kidan [83]. Generally, what has been stressed is the effect of excessive economic activity on imports and on the composition of investment and output. The idea of a trade–off is expressed by the "restraint policy" of the Israel government for 1965, [e.g., 158]. Here maximum growth was sacrificed in order to reduce inflationary pressure and the import surplus. The true cost of a temporary reduction in the import surplus, in terms of foregone output, was not properly estimated; this criticism was most forcefully presented by Beham and Kleiman [14]. Praeger [125] made a similar point, by presenting a crude esti-

[60] The admission by the Bank of Israel that the money supply must increase to prevent a decline in employment during cost–inflation has already been mentioned.

mate of a Phillips–Lipsey curve for Israel and concluding that a reduction in inflation was paid for by a tremendous increase in unemployment.

D. *Cost Inflation and Wage Policy*

Though most Israeli economists have been convinced that the local inflation has been more frequently of the demand type, the demand–pull versus cost–push controversy did not bypass Israel. Patinkin was probably the first to describe a cost–price spiral in Israel. He maintained that the July 1952–April 1953 period differed from the preceding inflationary period in that increases in the supply of money during this period were in response to rising prices, and not a cause of them [119].[51] He feared that Israel's institutional framework would perpetuate the inflationary process.

Two elements of cost–inflation have been singled out for analysis: devaluation and wage increases. Devaluation raises the price level significantly because of the relatively large volume of imports; but it does not create a cost–price spiral. This is done by nominal wage increases, resulting from an attempt either to raise real wages or to prevent a fall of real wages following price rises. Wage policy is primarily determined by the Histadrut—the General Federation of Workers in Israel—a mammoth organization encompassing virtually all wage earners (and many self–employed workers), dominated by the political leaders and parties who dominate the government. A basic part of wage policy has been a Cost–of–Living Allowance (COLA) based on changes in the Consumer Price Index (CPI). The nationally negotiated wage contracts include clauses specifying what change in the CPI warrants a wage allowance, the time period for which the computation is made, and what percentage of total wages is linked to the CLI. The details of the COLA are discussed in every Bank of Israel Report, in the chapter on wages, and have been summarized by Bahral [4].

The Histadrut's wage policy in general, and the role of the COLA in particular, were the core of Israel's inflationary problem, as analyzed by A. P. Lerner[52] [94, 95]. Lerner developed the concept of a Wage Standard,[53] where functional finance is not enough to prevent inflation because government monetary and fiscal policy must adjust to independently determined union wage policy. Lerner's recommendation was to set up a Wage Authority, which would set wages in accordance with expectations concerning changes in production and productivity, and thus prevent an inflationary spiral.

Although no Wage Authority was set up in Israel, the idea that the Histadrut should set its wage policy in accordance with national objectives was accepted, at least in theory. On several occasions, the Histadrut called for wage restraint, and even a wage freeze. The idea of linking wage increases to increases in productivity was adopted by government and Histadrut leadership in the early 1960's. Accordingly, a public commission was set up to measure projected changes in output per worker as a criterion for wage increases. So far, wage policy has not in fact been tied to the recommendations of this commission.

The most persistent exponent of the role of wage policy in Israel's inflationary ex-

[51] Though published in 1956, this article first appeared as a report to the Ministry of Finance in 1953.

[52] Abba Lerner is included among "Israeli" economists, because he spent several years in Israel as an adviser and a teacher at the Hebrew University. He has privately admitted that he "accepted" cost inflation in Israel. Certainly his writings on the subject in the daily press in Israel did much to stir up controversy and influence opinion.

[53] Lerner's Wage Standard is very similar to what Hicks has called a Labor Standard: J. R. Hicks "Economic Foundations of Wage Policy," The *Economic Journal*, September 1955.

perience has been Riemer. In several articles, most fully in [129] he has presented his version of the inflationary spiral, which is essentially as follows: The large influx of immigrants increased the supply of labor, absolutely and relative to other factors of production. The Histadrut, interested in the welfare of the *veteran population*,[54] prevented the required fall in real wages—in fact, real wages (nominal wages deflated by the CPI) doubled between the end of 1947 and the end of 1954. The marginal product of labor could have been prevented from falling if capital per worker had been increased sufficiently; but instead it declined during this period. The result was overt and disguised unemployment equalling 40 per cent of the labor force in 1954. The government tries to decrease unemployment by increasing the money supply, but since this does not lower real wages (from either the worker's or the employer's point of view) it can only result in demand inflation. This inflationary pressure can decrease unemployment at the expense of the balance of payments, thus monetary and wage policy act to shift disequilibrium between unemployment and the balance of payments.

Riemer's analysis has had less influence than it deserves—probably because of his accusatory style of writing and rather flamboyant use of figures. For example, though all agree that there may be considerable disguised unemployment in Israel, particularly in the public services,[55] no one else accepts his estimates; his total unemployment figure for 1954 assumes a normal labor force participation rate of 60 per cent, compared with an actual participation rate of about 38 per cent. Bahral's study of the impact of mass immigration on industrial wages [4] shows that though institutional factors delayed somewhat the expected market-induced changes in real factor prices, wages did in fact behave, with a lag, in accordance with market forces: while real wages, from the point of view of the worker, did increase, unskilled labor became cheaper relative to other factors and relative to products. Similarly, Gaathon's estimates of capital stock [51, 54] have contradicted Riemer's rough estimates: though there may have been some decline in capital per worker in the early years, capital per employed worker increased in 1950–55 at an annual rate of over 10 per cent, a higher rate than that achieved in later periods. Though these more thorough empirical studies undermine much of Riemer's analysis of the first decade, his analysis of the relationship between wage and monetary policy and the balance of payments, repeated in later articles [e.g., 130], remains of interest.

The most recent contribution to the cost inflation approach appears in the econometric model of the economy constructed by Evans [42]. Evans did not find that demand factors were statistically significant for explaining changes in prices. On the other hand, wages is a significant variable.[56]

The problem of identifying cost inflation has elicited some comment. Kochav [88] suggested that *ex post* cost inflation can be inferred if wages per worker have increased more than product per worker; he did not, however, describe his implicit

[54] At this point Riemer departs from Lerner's Wages Standard.

[55] This subject is dealt with by Ofer [115].

[56] His estimated equation for changes of prices in manufacturing and in agriculture show price changes as linear functions of changes in wages and productivity. The coefficient of wage changes is larger than the coefficient of productivity changes; this implies that had wages been geared to increases in output per worker—a policy officially adopted in the early 1960's—wage policy would still have been inflationary. However, the differences in the coefficients may be too small to be significant.

model. Klinov–Malul [65, pp. 274–276] has also suggested using changes in the share of wages in national income as a possible indication of cost or demand inflation. She explicitly assumes a Cobb–Douglas production function for the economy; i.e., she assumes that the share of wages should be constant in equilibrium. She furthermore assumes that inflation is the only cause for divergence from equilibrium. Thus, under cost inflation wages will lead, and under demand inflation wages will lag. Looking at the Israeli data, Klinov–Malul finds that a declining wage share between 1955–61 is consistent with demand inflation, and a rising wage share in 1964 and 1965 indicates cost inflation. Unfortunately, this test also suggests demand inflation for the 1952–55 period, a period which is generally recognized to have been a classic example of cost inflation. Klinov–Malul herself points out that Bruno [23] has explained the declining wage share as the result of market forces correcting disequilibrium in the early 1950's, when the price of labor exceeded the marginal product of labor; these findings are consistent with Bahral's analysis [4].

Another test employed by Klinov–Malul concerns the connection between the change in unemployment and the rise in average wages. In 1952–53 and again in 1965, she finds that unemployment and wages both rose, suggesting cost inflation, whereas in 1955–64 wages rose while unemployment fell.

In his estimates of various multipliers, Evans stresses the inflationary effects of wage increases. He finds that an increase in wages raises prices, which again raises wages; however, real wages increase. Lerner had pointed out how, in some cases, the COLA increased wages more than the increase in prices, and many have discussed the fact that exemption of the COLA from income tax increases real wages when prices rise. But Evans goes further: he stresses the fact that increases in wages raise some prices, but not all prices; assuming no devaluation, import prices do not rise, and prices of government services lag. Therefore, real wages and disposable income increase. He derives an upward sloping IS curve. Thus, wage increases lead to increases in real demand and employment, which lead to larger increases in wages and demand; the situation becomes explosive as full–employment is reached, and calls for government intervention.

Evans, Lerner and a special committee set up to study the workings of COLA[57] [173] all agree that it is intrinsically inflationary in cost inflation situations, given Israel's institutional system. However, the Committee pointed out that the COLA cannot initiate inflation, nor does it have any inflationary influence in a period of demand inflation except when, because of lags, the COLA adjustment of wages is made *after* the market has already raised wages.

The relationship between wages and prices has also been examined empirically by Kleiman and Ophir [84]. Using quarterly data, they constructed regressions for two periods, 1955–59 and 1960–65. Their regressions show that during the first period, two–fifths of the change in industrial prices is explained by changes in wages; taking into account the level of unemployment as a variable leads to a puzzling result: there is a positive correlation between unemployment and price increases. For the second period, they found that wage increases explain a smaller part of the rise in prices, but unemployment (and even more strongly, changes in un-

[57] This committee reached a consensus on all points (except one which really concerned an interpretation of its terms of reference), though made up of four economists representing different interests: B. Baer, of the Manufacturers' Association, D. Golomb, of the Histadrut, and E. Kleiman and D. Levhari of the Hebrew University.

employment) are negatively correlated with price increases. Taking the 1955–65 period as a whole, using quarterly industrial wages, they found that changes in the level of unemployment explain changes in wages better than the level of unemployment does (together with the CPI 45 per cent). Evans, using annual data, found that the unemployment level was not a major factor in explaining wage increases: his equations show that average annual increases are only slightly higher at a 10 per cent level of unemployment than at a 4 per cent level of unemployment.

E. *Linkage Schemes*

Linkage of wages to the CPI, by the COLA, is an arrangement that covers all wage and salary earners. During the 1950's, linkage of internal loans, investments and savings had gradually developed and was applied widely—the linkage being to the CPI, to the foreign exchange rate, or to both. In fact, the extensive use of linkage has been a main factor inducing government to worry about price rises, more than about the causes of price rises. There has been considerable discussion of wage linkage and its part in the inflationary process, but much less of the role of linkage as a hedge against the effects of inflation.

The Committee which examined the effects of the COLA [173] has shown that in a period of demand inflation the system is an excellent way to restore, without labor disputes, an income distribution (between capital and labor) that has been distorted by price rises. The less–than–perfect existing system, within Israel's less–than–perfect competitive labor market, the Committee believes, greatly reduced friction between unions and management during Israel's inflationary experience.

The Committee stresses that much misunderstanding arises from a confusion between two different objectives of the COLA system. One, already mentioned, is the preservation (or more exactly, restoration) of the income distribution between labor and capital during inflation. The second, in fact that most commonly thought of in connection with the system, is to preserve the real value of wages. This objective appears obviously desirable when prices are rising. However, if the rise in prices is due to a conscious attempt to lower real income, for example, by devaluation or indirect taxation, should there be an automatic system to negate it? The attempt to do so will lead to unemployment or cost inflation.[58] Automatic compensation for such increases in prices denies the government its right to devalue the currency, or to lower disposable income by indirect taxation.[59] This point has also been forforcefully made by Morag [109, 110 and 111, p. 170].

A misconception which is widely held in Israel is that the lag in payment of the COLA means that workers are not fully compensated for price increases.[60] The Committee pointed out that this is not true: when prices rise at a constant rate, wages rise—with a lag—at the same rate; when prices rise at a decreasing rate, wages rise (again, with a lag) at a faster rate, but when prices rise at an accelerated rate, wages rise more slowly. In all cases, an exact COLA system would exactly equate the index of wages to the index of prices,

[58] In this connection, it is worth citing Kohav's [88] contention that cost–inflation initiated by indirect taxes is misnamed, because government expenditures are planned independently of revenue, so that indirect taxes are alternatives to a government deficit, and are thus an "alternative to demand inflation."

[59] The Committee suggested excluding from the CPI index changes in import prices, but stopped short (except for a minority opinion by Baer) of suggesting the exclusion of indirect taxes, because of differences of opinion as regards terms of reference.

[60] Workers are of course not fully compensated if the COLA covers less than the total wage, as is usually the case for the higher wage recipients.

with a one period lag. However, attention should be paid to the periods compared; there is an initial loss, thereafter compensation is adequate or more.

Escalator clauses for loans were first introduced in 1949, but only gradually during the 1950's did they become widespread. The linkage of loans has been surveyed by Sarnat [133, pp. 71–74] and Morag [112]. Sarnat has pointed out that private rather than government initiative played a major role in the development of the system. He also stressed the role of linkage in the development of the bond market [134, p. 107], and in the development of savings through provident pension funds [133]. The effective rate of interest or return, taking into account all nominal charges and linkage to the foreign exchange rate, or to the CPI, has been estimated for loans by Ben–Shahar [16], for investments of retirement funds by Sarnat [133] and for bonds by Ben–Shahar and Sarnat together [17].

These data are relevant to a problem frequently discussed: To what should loans be linked, the foreign exchange rate or the CPI? In an early discussion of this question designed as a guide to holders of bonds who had to choose between the two kinds of linkage, Riemer [128] pointed out that linkage to the CPI gives security, if the index can be considered an accurate measure of the price level of consumer goods for middle–class investors, whereas linkage to the rate on the dollar answers a quest for speculation. On the other hand, investors who had foreign debts linked to the dollar would find security in linkage to the exchange rate.[61] Doroth–Dusterwald [35] also stressed the danger to firms and even to the government of having liabilities linked to the dollar when assets are not linked at all, or are linked to some other measure. Ben–Shahar and Sarnat show that for the 1958–64 period (a formal devaluation was declared only in February 1962) dollar–linked bonds gave a higher rate of return than those linked partly to the dollar and partly to the CPI. Bonds linked only to the CPI gave a lower return, and unlinked bonds only a very slight return.

The Bank of Israel [152] argued that linkage to the foreign exchange rate was undesirable because of the sporadic and extreme nature of changes in the exchange rate. Morag, a strong proponent of escalator clauses in general, also objected to foreign linkage [112, p. 149] on the grounds that it has implications of an "abdication of the local currency," that it frustrates devaluation (even when needed), and that the linkage charges actually paid or received would vary very much and depend on how near to the date of devaluation the obligations were incurred. This last point is well brought out by the Ben–Shahar–Sarnat data. After the 1962 devaluation, the government tried to discourage linkage to the foreign exchange rate. In recent years, all linkage of loans has been considerably reduced, by introduction of an optional four per cent "insurance" against price rises surcharge instead of linkage to the CPI.

As regards linkage in general, Morag has been the most articulate spokesman for linking everything to some price index, and thus making the entire economy inflation–proof. Morag saw the evils of inflation not in the mere rise in the price level, but in its effect on the growth and distribution of real product. These adverse affects could be avoided by general escalator clauses [111, pp. 154–173]. To those, such as Horowitz [76], who feel that the major evil of inflation lies in its effects

[61] When these holders of the bonds discussed, issued by the Palestine Electric Corporation in 1955, made their decision, 80 per cent opted for linkage to the consumer price index; Sarnat maintains that this choice reflects the large institutional (retirement funds) demand for these bonds [133, p. 41].

on the balance of payments, Morag's solution is little help; they would reject the implied necessity of reliance on frequent devaluation to solve the balance of payments deficit. Thus, they suggest restraining inflation rather than using linkage to mitigate its effects.

When Morag presented his suggestion at the 3rd Rehovot Conference in 1965, two of the discussants who argued against it, Richard Goode and Harry G. Johnson, raised the point that such an inflation-proof scheme would eliminate the use of inflation as a method of redistributing income and wealth. This of course implies that one can regard inflation as having some desirable (or at least useful) effects. As mentioned above, the only one to defend inflation in Israel—and then only limited inflation—has been Schweizer [138], who stressed that it stimulates investment. Though this view is in the extreme minority, the Israel experience would suggest that it should be more popular: inflation has been used for almost twenty years as a means of increasing investment, creating forced savings, and redistributing income. This observation leads us to a final comment; though there have been many reflections and oracular statements about inflation in Israel, there has been no serious empirical investigation or even serious discussion of whether the inflationary policy was on the whole justified, or whether Israel's economic achievements have been attained without inflation. In this respect, Israeli economists share a failing with economists the world over.

V. *Agricultural Policy*

Agriculture in Israel is much more than a branch of the economy; it is a cherished way of life. The concept of socialist-Zionism was rooted in the ideal of restoring the Jews of the diaspora to land and to labor, in Palestine. Any history of the modern restoration of Jewish Palestine would have to lay great stress on the development of agricultural settlements, particularly the *moshav* (cooperative village) and *kibbutz* (collective village). The development of these forms of settlement between 1900 and 1947 was, to a large extent, formed by the sociological, political and economic conditions prevailing at the time; the great success of these settlements during the pre-State period did much to determine the attitude to agriculture and agricultural settlement after 1948.

Three factors—the influx of new immigrants, the cutting off of Israel from traditional sources of food, and the sudden availability of land vacated by Arabs—combined to make massive agricultural settlement an obviously wise decision. Within three years hundreds of settlements were set up, and the cultivated area doubled; the following years were marked by tremendous investment in irrigation. Though agricultural development was slowed down after the first few years, agriculture has remained subject to a degree of public control not usually found in non-Communist countires.[62] The overwhelming public ownership of land (93 per cent), the continuous vesting of responsibilities for colonization policy in the hands of the Jewish Agency for Palestine (in cooperation with the Ministry of Agriculture), and the affiliation of the collective and cooperative villages with the Histadrut, all combine to explain this. Even citriculture—where the private sector is strongest—is subject to control through a Citrus Marketing Board.

The debates concerning agricultural policy have covered many topics: planning, colonization, production, marketing and pricing of goods and factors. In general, it may be said that the debates mainly relate to the application of economic

[62] A survey of agricultural development in Israel is contained in Mundlak [114, Ch. 2].

considerations to agricultural policy, and the way various instruments of policy take them into account. Three examples of policy discussion are: farm structure, the pricing of water, and agricultural planning.

A. *Farm Structure*

Though historically private farming preceded the *kibbutz* and *moshav*, the latter has gained prominence. The reluctance of most new immigrants to adopt the more communal life of the *kibbutz* led to placing greater stress on the *moshav* in colonization policy after 1948. During the pre-State period, political restrictions on land purchase and emphasis on "absorptive capacity" in immigration policy led to the development of a farm which accented mixed farming: great labor intensity, diversity of crops, and concentration on high-value foodstuffs. The mass colonization of the early post-1948 years tended to reproduce the type of farm which had proven so successful in the past. But the basic economic conditions had changed: land was now abundant, and in a short period of time it became apparent that water was the scarce factor.[63] This fundamental change was emphasized by a committee examining Israel's agriculture in 1953/54 [165], whose report reflected the tremendous contribution to the introduction of economic considerations into agricultural policy of Marion Clawson, then a member of the Economic Advisory Staff. Moreover, by 1954 the supply of the high-value mixed-farming products was outrunning demand; whereas previously price control was used to keep prices down, now controls and subsidies were widely used to protect farm income. Albert G Black, the FAO representative in Israel—who shares with Clawson the credit for influencing local thought—attacked the form of the *moshav* [18]. He pointed out that the structure of the farm could not support efficient agriculture, and recommended a switch to field-crops, farmed on large tracts of land. A major defect of Black's recommendations, in the sense that it made them unacceptable, was his attack on the basic concept of the family farm. Weitz, a senior official (later director) of the Jewish Agency Colonization Department, adopted Black's ideas but tried to adapt them to the standard concept of the family farm and *moshav*. He argued [53] that the family farm could be efficient if more land was allocated per settler. Thus, each family would be guaranteed some income from higher-value crops and livestock near the house, and also work efficiently on field crops if a portion of each family's land formed part of an extensive belt farmed cooperatively. Weitz described different types of farms, suitable for different regions of the country. In time, greater diversity in farms was both planned and implemented. Weitz and Rokach have done much to publicize the Israeli experience in rural planning, based on the idea of the family farm, and its suitability as a model for developing countries, e.g. [144, 145].

The economic planning of farm settlements benefitted considerably from studies of established family farms, initiated by Kaddar with the Falk Project for Economic Research in Israel in 1954, taken over by Lowe and the Ministry of Agriculture, and later by Mundlak, who published the most detailed results of these studies [113]. Based on a sample of established farms in cooperative villages, these studies surveyed developments in output and the response to price changes, and estimated production functions. Mundlak has shown how the administrative allocation of land and water led farmers to respond to administered market prices by concentrating on cattle and poultry pro-

[63] The importance of the changes in these factors was pointed out as early as 1951, for example by Samuel [132].

ducts, and expanding output by use of large inputs. As a consequence, the marginal product of a unit of land and water was found to be zero, because a shift from cattle and poultry (low land–water users) into crops or forage did not add to the value of production. A more far–reaching conclusion was that newer farms could not follow the same path to reach a high standard of living,—comparable to that of urban workers, because of the saturation of the market. Consequently, a switch to larger farms and other products was recommended. That the existing subsidization of products went primarily to increase the income of established farms and led to great interfarm income inequality was also the conclusion of a committee studying the farm situation in 1959 [172].

B. *The Pricing of Water*

The attitude toward agriculture in general is clearly evident in the never–ending discussion of the pricing of water. In Israel, most of the cultivable land is in the south, where rainfall is slight, whereas the main sources of water are in the north. The concept of a country–wide water system was envisaged during the pre–State period; but only after independence was achieved could state planning and implementation of the water network be undertaken. With the completion of the Jordan River Project, all the major regional water schemes are interconnected, making possible not only transfers of water from the north to the south, but also balancing of all regional surpluses and shortages. These projects turned out to be quite expensive; thus water costs are high by international comparisons. Moreover, over the years the estimates of total water resources became increasingly less optimistic; at present, it is clear that conventional sources of water (i.e. exclusive of desalination) can irrigate no more than half the cultivable land. Thus the questions where to use the water and who should pay for it became crucial, the more so since a large part of the total water supply is provided by a public water company.

The most extreme position is that water should be free, like air. As soon as water was recognized to be a scarce factor, even the more optimistic water–planners took a less extreme position. Blass, for many years director of water planning and the most optimistic regarding available water resources, defended the idea that water must bear a price to the consumer to prevent wastage—by which he means simply using too much, or not turning off taps [19]. He also recognized that a unitary national price would be neither economic nor just. He differentiated between the real cost of production of water, which he understated by taking a 3 per cent rate of interest as "suitable" for public projects, and the price which should be charged to farmers. Farmers could not be expected to pay the full cost of water because it is too expensive in distant areas and too cheap near the water sources. Kariv, a senior official of the water supply company, came out against the double subsidization of water, once by underestimating costs, and again by selling to farmers below estimated costs [80]. He pointed out the misallocation of resources arising from such pricing. Though advocating estimating costs more realistically, he did not clearly specify how prices to the farmers should be determined.

More generally, the following argument must be answered: In Israel, development of under–populated areas has been given national priority, and fostered by direct transfer of immigrants to agricultural settlements. Must they bear the price of high water costs? The system used to subsidize water implies that in part water is considered a social good, part of whose cost is covered by the government. At a

symposium on the economics of water held at the Hebrew University in 1965 [168], one suggestion raised was to consider the availability of water in each area as a national burden; farmers would pay only marginal costs, these being charged in order to induce them to use water economically in their choice of crops. The exclusion of fixed costs in estimating the price of water is common policy, but the exact establishment of marginal costs and marginal cost pricing has not been done.

As a basis for more rational pricing of water, two detailed studies have been carried out. Yaron has estimated the agricultural demand for water. He used a double approach: sample studies for certain areas to determine the marginal product of water in the past, and a linear programming model to estimate more generally the optimal change in production plans and farm income as a result of changes in the amount of water [147, 148]. Ophir has made short–run cost estimates of the supply of water, and taking demand estimates as parameters, he has used a linear programming model to estimate long–run supply curves. Examining fifty projects for the development of new water sources, the model shows the optimum sequence of water development. While the final results have not yet been published, the approach has been described in [117] and some preliminary results presented in [116].

Though much progress has been made in estimating the economic variables necessary for rational use of water resources, less has been done to show how price policy can be used within Israel's institutional constraints. The basic problem has been clearly formulated in a symposium carried out in 1965 at the Hebrew University Department of Agricultural Economics. Replying to a suggestion by Mundlak that *moshav* farmers should be allowed to trade water rights, Gvati (later Minister of Agriculture) stated that though it was possible that such a policy would increase economic efficiency, he was certain that it would undermine the social–economic structure of the *moshav.*

C. *Agricultural Planning*

The development of agricultural planning in Israel reflects the progress of economic thought concerning agriculture. Planning is undertaken jointly by the Ministry of Agriculture (responsible for agricultural policy) and the Jewish Agency Colonization Department (responsible for new agricultural settlements). There has been a series of four, five and even seven year plans; it may be said that planning is continuous.

The first plans [e.g., 167] were mainly physical, not economic. They concentrated on maximization of food production, and on economizing on foreign exchange, and assumed that 25 per cent of the population should subsist from agriculture. The early seven–year plans were not very accurate forecasts; comparing the original 1963 plan for 1959/60 with actual production in 1959/60, Lowe has shown [99] that water usage was grossly over–estimated and non–irrigated cultivated land grossly under–estimated; production developed more than planned in poultry, beef and vegetables, and much less than expected in field crops.

Gradually, economic considerations were introduced; however, no single clear cut criterion is evident. Rubner [131, p. 111] summarized the following often contradictory criteria used for planning: maximum income of farmers, maximum production of calories, maximum foreign exchange earnings, maximum value of output per acre, maximum value of output per unit of water, and maximum employment per unit of land.

Mundlak supervised a massive study of demand and supply of agricultural com-

modities, making forecasts for 1965 and 1975 [114]. The approach of the study was as follows: (1) Assuming 1960 prices to remain unchanged, demand was estimated on the basis of projections of income and studies of income elasticity (mainly Liviatan's [98]). (2) Supply was estimated on the basis of past trends and anticipated changes in productivity and wages, again under the assumption of constant relative prices. (3) Finally, prices were allowed to vary, and equilibrium prices were estimated taking into account cross–elasticities of demand. Mundlak's main conclusion [Vol. I, Ch. 1] was that only relatively small price adjustment could be anticipated, therefore the dislocation involved in transferring from intervention to reliance on a free price mechanism would be small. Though he did not suggest a detailed agricultural policy, he advocated a new approach, to wit: reliance on the price mechanism in both the product and the factor markets.

In fact, his approach has not been adopted. But the detailed knowledge carefully accumulated on production functions, income and price elasticities have been taken into account in recent planning. The joint effort of many economists, including those of the Hebrew University Department of Agricultural Economics, the recent five–year plan [166] departs from its predecessors both in its dove–tailing of national, regional and local plans, and in its incorporation of policy recommendations for development of internationally tradeable goods (evaluated on the basis of cost per dollar), fluctuating –price local market products, and other domestic goods. Conceding that pricing policy in the past was mainly income–oriented (and not too successfully at that) and ignored allocation effects, Yaacobi [146] maintains that the new plans call for much more balanced use of price and other intervention policies, with gradual elimination of distorting subsidies. How this will actually develop remains to be seen.

References*

1. Z. Abramowitz, "The Balance Sheet of the Israel Economy," *EQ*, Sept., 1953, *1–2*, 36–41, H.
2. A. Arian, "On Contestation of the Quantity Theory," *EQ*, May 1965, *45–46*, 174–175, H.
3. U. Bahral, *The Real Rate of the Dollar in the Economy of Israe* . Jerusalem, 1956, H.
4. ———, *The Effect of Mass Immigration on Wages in Israel.* Jerusalem, 1965, H/E.
5. H. Barkai, "Consumption of Edible Oil in Israel and Supply of Local Materials for Their Production," Eliezer Kaplan School of Economics and Social Sciences, *Studies in Economics*, No. 1. Jerusalem, 1956, H.
6. ——— and M. Michaely, "The New Economic Policy After One Year," *EQ*, March 1963, *37–38*, 23–40, H.
7. ——— and ———, "More About the New Economic Policy," *EQ*, Aug. 1963, *39*, 210–232, H.
8. J. Baruh, "Import Taxes and Export Subsidies in Israel, 1955–1961," *BIB*, March 1963, *18*, 48–70, H/E.
9. I. Bar-Yosef, "Credit Policy in Israel," *E Q*, June 1954, *3*, 187–195, H.
10. ———, "The Structure of Interest During Inflation," *EQ*, Jan. 1960, *25–26*, 86–97, H.
11. ———, "Israel Monetary Policy in the Test of Reality," *EQ*, Nov. 1962, *36*, 351–61, H.
12. ———, "Some Observations on a Recommendation to Tighten Monetary Policy," *Quarterly Banking Review*, Dec. 1962, *7*, 49–58, H.
13. M. Beham, *Monetary Aspects of the 1962 Devaluation.* Jerusalem, 1968, H/E

* References marked H are in Hebrew only, E in English, and H/E appeared in both Hebrew and English editions. The most frequently referred to journals are *The Economic Quarterly* and the *Bank of Israel Bulletin*, abbreviated to *EQ* and *BIB*, respectively.

14. ——— AND E. KLEIMAN, "The Cost of Recession," *Quarterly Banking Review*, June 1968, *29*, H.
15. H. BEN-SHAHAR, "Means of Payment and the Balance of Payments," *EQ*, March 1962, *33–34*, 54–58, H.
16. ———, *Interest Rates and the Cost of Capital in Israel, 1950–1962*. Basel and Tubingen, 1965, E.
17. ——— AND M. J. SARNAT, "New Issues of Profitability of Investment in Common Stock 1959–1964," *BIB*, April 1960, *25*, 3–31, H/E.
18. A. G. BLACK, "Reflections Upon Israel's Recent Agricultural Development and its Relationship to General Development," in *The Challenge of Development*, Jerusalem, 1958, E.
19. S. BLASS, "The Problem of the Price of Water for Agriculture," *EQ*, May 1954, *4*, 364–372, H.
20. A. BONNÉ, "Productive and Wasteful Capital," *EQ*, Sept. 1953, *1–2*, 19–25, H.
21. A. BOROKHOV, "Land Prices and Inflation," *EQ*, Jan. 1965, *44*, 348–354, H.
22. M. BRUNO, *Interdependence, Resource Use and Structural Change in Israel*. Jerusalem, 1962, H/E.
23. ———, "Estimation of Factor Contribution to Growth Under Structural Disequilibrium," *International Economic Review*, Feb. 1968, *9*, 49–62, E.
24. ———, "The Optimal Selection of Export-Promoting and Import-Substitution Projects," in United Nations, *Planning the External Sector: Techniques, Problems and Policy*, 1967, E.
25. ———, "A Programming Model for Israel," in I. Adelman and E. Thorbecke, ed., *The Theory and Design of Economic Development*, Baltimore, 1966 E.
26. ———, "Optimum Patterns of Trade and Development," *Review of Economics and Statistics*, Nov. 1967, *49*, 545–554, E.
27. ———, M. FRAENEKL AND C. DOUGHERTY, "Dynamic Input-Output, Trade and Development," presented at the Fourth International Conference on Input-Output Techniques, Jan. 1968; forthcoming in *Essays in Honor of W. W. Leontieff*, E.
28. H. CATS, "A Study of Value Added in Terms of International Prices," in Falk Project for Economic Research in Israel, *Third Annual Report 1956*, Jerusalem, 1957, 37–40, H/E.
29. ———, "Labor Productivity and Value Added," *EQ*, Feb. 1961, *29–30*, 64–75, H.
30. H. B. CHENERY AND M. BRUNO, "Development Alternatives in An Open Economy: The Case of Israel," *Econ. Jour.*, March 1962, *72*, 79–103, E.
31. Z. CITRON AND A. KESSLER, *Investments in Manufacturing Made Through the Investment Center*, Falk Project for Economic Research in Israel, Research Paper 2. Jerusalem, 1958, H/E.
32. M. CLAWSON, "General Survey of Agriculture in Israel," Economic Advisory Staff, 1953. Mimeograph, E.
33. D. CREAMER, *Israel's National Income 1950–1954*. Jerusalem, 1957, H/E.
34. F. DAVIDOV, "The Liberalization of Imports—February 1962 to May 1965," *BIB*, June 1966, *26*, 49–64, H/E.
35. A. K. DOROTH-DUSTERWALD, "Linkage of Loans to the Real Value of the Currency," *EQ*, Oct. 1954, *5–6*, 68–74, H.
36. E. DOVRAT, "Considerations for Economic Policy," *EQ*, Dec. 1968, *59–60*, 189–201, H.
37. ———, N. WAHL AND M. TAMARI, "The Financial Structure of Israel's Industrial Companies," *BIB*, Jan. 1963, *17*, 39–93, H/E.
38. F. DULBERG (Ginor), "Consumption, Production and Economic Equilibrium," *EQ*, Oct. 1954, *5–6*, 19–29, H.
39. ———, "The Trade Balance and Economic Progress," *EQ*, June 1955, *8*, 368–378, H.
40. ———, "Economic Growth Without Inflation," *EQ*, July 1956, *12*, 325–341, H.
41. ———, "Exports and Economic Independence," *EQ*, March 1957, *15*, 193–210, H.
42. M. K. EVANS, "An Econometric Model

of Part of the Israel Economy," Wharton School Discussion Paper No. 86, 1968, E.
43. H. FRUMKIN, "Economic Independence or Economic Strengthening," *EQ*, June 1961, *31*, 299–302, H.
44. A. L. GAATHON, "Foundations for Economic Independence," *EQ*, Oct. 1956, *13–14*, 24–43, H.
45. ———, "Increased Immigration and the Balance of Payments," *EQ*, Nov. 1957, *17–18*, 22–26, H.
46. ———, *Survey of Israel's Economy in 1951*. Jerusalem, 1959, H/E.
47. ———, "A Note on the Treatment of Multiple Exchange Rates in National Accounts," *Rev. of Econ. Stat.*, Aug. 1960, *13*, 326–328, E.
48. ———, *Capital Stock, Employment and Output in Israel 1950–1959*. Jerusalem, 1961, H/E.
49. ———, "Economic Planning in Israel: History and Problems," in Y. Ronen, ed. *Israel Economy: Theory and Practice*, Tel-Aviv, 1964, H.
50. ———, "Official Rate of Exchange or Effective Rate of Exchange in the National Accounts," *EQ*, Feb. 1967, *52*, 378–380, H.
51. ———, *Economic Productivity in Israel 1950–1965*. Jerusalem and New York, forthcoming.
52. A. GAFNI, N. HALEVI AND G. HANOCH, *Israel's Tariff Structure and Functions*, Falk Project for Economic Research in Israel, Research Paper 3. Jerusalem, 1958, H.
53. ———, "Classification of Tariffs by Functions," *KYKLOS*, 1963, *16*, 302–320, E.
54. M. GILBOA, "Criteria of Rentability, Production and Investment," *EQ*, Nov. 1957, *17–18*, 73–92, H.
55. F. GINOR, *Uses of Agricultural Surpluses*. Jerusalem, 1963, H/E.
56. T. GOLDBERGER (Ophir), *Criteria for Protective Tariff Policy in Israel*. Jerusalem, 1956. Mimeograph, H.
57. D. GOLOMB, "The Import Surplus—Cause or Effect," *EQ*, Feb. 1961, *29–30*, 150–152, H.
58. ———, "On the Causes of Inflation," *EQ*, Nov. 1962, *36*, 150–152, H.
59. M. GOREN, "Considerations of Investment Policy in Manufacturing Enterprises," *EQ*, June 1960, *27*, 277–285, H.
60. S. GOTTLIEB, "Government Subsidies in Israel: 1954 and 1955," in Falk Project for Economic Research in Israel, *Third Annual Report 1956*, 26–28, Jerusalem, 1957, H/E.
61. N. GROSS, *Inflation and Economic Policy in Israel: The First Stage, 1949–1950*. Jerusalem, 1952. Mimeograph, H.
62. A. L. GRUENBAUM (Gaathon), *Four Year Development Plan 1950–1953*. Jerusalem, 1950, H. (abridged English edition)
63. I. GUELFAT, "Economic Planning and Development," *EQ*, Dec. 1958, *21–22*, 110–123, H.
64. N. HALEVI, "Exchange Control in Israel," in P. Uri, ed., *Studies on the Israel Economy*, Paris, forthcoming.
65. ———, AND R. KLINOV-MALUL, *The Economic Development of Israel*. Jerusalem and New York, 1968, E/H.
66. A. HALPERIN, "The Free Credit Market," *EQ*, June 1956, *12*, 410–419, H.
67. ———, "Monetary Policy: Forecasts and Reality," *EQ*, Aug. 1963, *39*, 233–247, H.
68. ———, "On the Failure of the Theory of Internal Infusion," *EQ*, Jan. 1965, *44*, 263–269, H.
69. M. HETH, *Banking Institutions in Israel*. Jerusalem, 1966, H/E.
70. ———, *Flow of Funds in the Israel Economy 1959–1966*. Jerusalem, 1968, H.
71. D. HOROWITZ, "Monetary Policy in Israel," *EQ*, May 1954, *4*, 295–305, H.
72. ———, "Summary of Report by the Governor of the Bank of Israel on the Increase in the Means of Payment," *BIB*, Dec. 1955, *2*, 3–6, H/E.
73. ———, "Report by the Governor of the Bank of Israel on the Increase in the Means of Payment During the Period 31 Aug. 1955 to 31 Aug. 1956," *BIB*, March 1957, *4*, 33–48, H/E.

74. ———, "Report by the Governor of the Bank of Israel on the Increase in the Means of Payment for the Period 31 Aug. 1956 to 30 May 1957," *BIB*, Oct. 1957, *6*, 55–78, H/E.
75. ———, "On the Threshold of the Second Decade," *EQ*, April 1958, *19*, 217–229, H.
76. ———, "Monetary Policy in Israel," in Y. Ronen, ed., *Israel Economy: Theory and Practice*, Tel–Aviv, 1964, H.
77. ———, "Economic Growth and Economic Independence," *EQ*, June 1966, *49–50*, 3–22, H.
78. ———, "Inter-relations Between Balance of Payments, Prices and Employment," *EQ*, Oct. 1966, *51*, 188–190, H.
79. M. Kalecki, *Report on Main Current Economic Problems of Israel*. Tel-Aviv, 1950. Mimeograph, E.
80. Z. Kariv, "The Water Economy—Problems of Pricing and Financing," in Y. Ronen, ed., *Israel Economy: Theory and Practice*, Tel–Aviv, 1964, H.
81. D. Katz, "The Causes of the Rise of Prices in 1960/61," *EQ*, Nov. 1962, *36*, 328–338, H.
82. A. Kerem, "The Economic Significance of the Interest Rate," *EQ*, Oct. 1955, *9–10*, 45–50, H.
83. A. Kidan, "Production, Imports and Demand," *EQ*, Aug. 1958, *20*, 364–370, H.
84. E. Kleiman and T. Ophir, "The Development of Inflation in Israel," *BIB*, forthcoming, H/E.
85. M. Klibanski, "Investment of Foreign Capital," *EQ*, Aug. 1963, *39*, 256–264, H.
86. D. Kochav, "Credit Policy," *EQ*, Oct. 1955, *9–10*, 39–44, H.
87. ———, "Remarks on Economic Planning," *EQ*, May 1959, *23*, 244–246, H.
88. ———, "Monetary Developments in Israel," in Y. Ronen, ed., *Israel Economy: Theory and Practice*, Tel–Aviv, 1964, H.
89. M. Landsberger, "Changes in the Consumer Price Index and Different Income, Origin, and Family Size Groups, 1954–62," *BIB*, Nov. 1963, *19*, 64–74, H/E.
90. ———, "Windfall Income and Consumption: Comment," *Am. Econ. Rev.*, June 1966, *56*, 534–545, E.
91. ———, *The Effect of Windfall Income on Consumption and Composition of Savings and Its Implications for the Income Concept in the Consumption Function*. Bank of Israel, forthcoming, H/E.
92. E. Lehman, "On Credit Policy," *EQ*, March 1963, *37–38*, 57–60, H.
93. A. P. Lerner, "Stabilization and Economic Independence," *EQ*, Oct. 1954, *5–6*, 3–11, H.
94. ———, "Inflation: Will it Wreck the Israel Economy in 1956?" Jerusalem, 1955. Mimeograph, H/E.
95. ———, "The Histadrut and the Israel Economy," *Midstream*, Autumn 1957, E.
96. E. Levy and Others, *Israel's Income and Expenditure, 1950–1962*. Jerusalem, 1964, H/E.
97. N. Liviatan, "Tests of the Permanent-Income Hypothesis Based on a Reinterview Savings Survey," in *Measurement in Economics: Studies in Mathematical Economics and Econometrics in Memory of Yehuda Grunfeld*, Stanford, 1963.
98. ———, *Consumption Patterns in Israel*. Jerusalem, 1964, H/E.
99. Y. Lowe, "Problems of Agricultural Production and Its Economic Expansion," in Y. Ronen, ed., *Israel Economy: Theory and Practice*, Tel–Aviv, 1964, H.
100. H. Lubell and Others, *Israel's National Expenditure: 1950–1954*. Jerusalem, 1958, H/E.
101. M. Mandelbaum, "Rentability Criteria in Industry," *EQ*, March 1963, *37–38*, 136–143, H.
102. ———, and M. Naveh, "Liberalization of Imports," *EQ*, Dec. 1963, *40*, 346–356, H.
103. M. Merhav, "Economic Planning and Industrialization," *EQ*, Dec. 1958, *21–22*, 30–51, H.

104. ———, "Some Economic Consequences of a Large-Scale Renewal of Immigration," *BIB*, Jan. 1963, *17*, 3–38, H/E.
105. M. Michaely, "Capital Imports into Israel: Source, Size and Meaning," *EQ*, June 1961, *31*, 217–241, H.
106. ———, *Foreign Trade and Capital Imports in Israel*. Tel-Aviv, 1963, H.
107. ———, "Foreign Exchange Rates in National Accounting," *Economica*, Aug. 1967, 289–297, E.
108. ———, *Israel's Foreign Exchange Rate System*. Jerusalem, 1968, H.
109. A. Morag, "Toward an Inflation-Proof Economy," *Am. Econ. Rev.*, March 1962, 177–184, E.
110. ———, "Tax on Value Added," *EQ*, March 1964, *41–42*, 67–73, H.
111. ———, *On Taxes and Inflation*. New York, 1965, E.
112. ———, "Value Linking of Loans in Israel," in D. Krivine, ed., *Fiscal and Monetary Problems in Developing States*, New York, 1967.
113. Y. Mundlak, *An Economic Analysis of Established Family Farms in Israel 1953–1958*. Jerusalem, 1964, H/E.
114. ———, and Others, *Long-term Projections of Supply and Demand for Agricultural Products in Israel*. 2 Vol. Jerusalem, 1964, H/E.
115. G. Ofer, *The Service Industries in a Developing Economy: Israel as a Case Study*. Jerusalem and New York, 1967, E.
116. T. Ophir, "The Cost of the Quality of Water," in *Development of Water Resources in Israel: Economic Aspects*, Jerusalem, 1965, H.
117. ———, "Optimal Development of Water Resources Over Time," *Israel Journal of Technology*, Oct. 1968, E.
118. D. Patinkin, "Credit and Interest," *EQ*, Oct. 1955, *9–10*, 9–10, H.
119. ———, "Monetary Developments in Israel," *Scripta Hierosolymitana*, 1956, *3*, 27–52, E.
120. ———, *The Israel Economy, The First Decade*. Jerusalem, 1960, H/E. (First issued in Falk Project for Economic Research in Israel, *Report 1957 and 1958*, 1959, H/E.)
121. ———, "Economic Advancement in Israel," *EQ*, Jan. 1960, *25–26*, 22–28, H.
122. ———, *On the Nature of the Monetary Mechanism*, Wicksell Lectures 1967, Stockholm, 1967, E.
123. D. Pines, "Criteria for Economic Independence," *EQ*, June 1961, *31*, 242–251, H.
124. ———, *Direct Export Premiums in Israel: 1952–1958*. Jerusalem, 1963, H.
125. J. Praeger, "Inflation and Unemployment—Is a Choice Inevitable?", *Quarterly Banking Review*, March 1968, 7, H.
126. S. Riemer, "Israel's New Economic Policy (NEP): The Background," *Public Finance*, 1954, *9*, 267–304, E. (Written under the pseudonym Saaki Eretz Ahuva.)
127. ———, "Israel's New Economic Policy in Action," *Public Finance*, 1954, *9*, 315–360, E. (Written under the pseudonym Saaki Eretz Ahuva.)
128. ———, *Index or Dollar?* Union Bank of Israel, 1955, E.
129. ———, "Inflation in Israel," *Public Finance*, 1957, *12*, 240–277, E.
130. ———, "The Devaluation of the Israel Pound," *KYKLOS*, 1962, *15*, 657–670, E.
131. A. Rubner, *The Economy of Israel*. London, 1960, E.
132. L. Samuel, "Agricultural Development During 1949–50 and Prevailing Tendencies," in E. Kolleck, ed., *The Israel Economic Annual 1951*, Jerusalem, 1951, E.
133. M. Sarnat, *Saving and Investment Through Retirement Funds in Israel*. Jerusalem, 1966, H/E.
134. ———, *The Development of the Securities Market in Israel*. Basel and Tubingen, 1966, E.
135. D. Sharon, "Bank Credit and the Level of Production," *EQ*, Feb. 1955, *7*, 247–253, H.
136. E. Scheffer, "Liquidity Rates and

Credit," *EQ*, Nov. 1962, *36*, 412–415, H.

137. A. STEINBERG, "Bill Brokerage in the Israel Economy," *BIB*, March 1968, *30*, 3–35, H/E.
138. A. SCHWEITZER, "Production and Credit," *EQ*, Oct. 1955, *9–10*, 33–38, H.
139. M. TAMARI, "Changes in the Financial Structure of Israel's Industrial Companies, 1956/57 to 1961/62, *BIB*, Nov. 1963, *19*, 14–63, H/E.
140. B. TOREN, *Criteria for Examining Investment Programs*. Jerusalem, 1957. Mimeograph, H.
141. N. WAHL, "Foreign Trade Accounts in Local Currency," *EQ*, Feb. 1967, *52*, 370–377, H.
142. Y. WEISS, *Price Control in Israel 1949–1963*. Jerusalem, 1964. Mimeograph, H.
143. R. WEITZ, *Our Way in Agriculture and Colonization*. Tel–Aviv, 1959, H.
144. ———, (General Editor), *Rural Planning in Developing Countries*. London, 1965, E.
145. ———, AND R. ROKACH, *Agricultural Development, Planning and Implementation* (*Israel Case Study*). Dordrecht and New York, 1968, E.
146. G. YAACOBI, "Five Year Plan for Development of Agriculture and Rural Areas 1966/67–1970/71," *EQ*, Feb. 1967, *52*, 392–401, H.
147. D. YARON, "Efficient Use of Water and Development of Water Sources," *EQ*, Aug. 1963, *39*, 288–296, H.
148. ———, "The Demand for Water in Israel Agriculture," in *Development of Water Resources in Israel: Economic Aspects*, Jerusalem, 1965, H.
149. S. ZAK, "On Credit Policy," *EQ*, Oct. 1954, *5–6*, 44–48, H.
150. F. ZUSSMAN, "A Development Plan for Agriculture," *EQ*, Sept. 1964, *43*, 259–276, H.
151. Z. ZUSSMAN, "The Limitations of Economic Planning in Israel," *EQ*, Oct. 1960, *28*, 337–343, H.
152. BANK OF ISRAEL, "Purchasing Power Guaranteed Securities in Israel," *BIB*, July 1962, *3*, 24–28, H/E.
153. ———, *Annual Report 1957*. Jerusalem, 1958, H/E.
154. ———, *Annual Report 1967*. Jerusalem, 1968, H/E.
155. ———, *German Reparations and Their Effect on the Israel Economy*. Tel–Aviv, 1965, H.
156. BANK OF ISRAEL RESEARCH DEPARTMENT, *Comments on the Report "Israel's Industrial Finances . . . A Second Look."* Jerusalem, 1961, H/E.
157. BANK OF ISRAEL RESEARCH DEPARTMENT AND MINISTRY OF FINANCE, "National Budget for the Fiscal Year 1959/60," *BIB*, Oct. 1959, *10*, 42–62, H/E.
158. BANK OF ISRAEL RESEARCH DEPARTMENT MINISTRY OF FINANCE AND ECONOMIC PLANNING AUTHORITY, "National Budget for the Fiscal Year 1965/66," *BIB*, March 1965, *23*, 63–92, H/E.
159. CONTINENTAL–ALLIED CO., INC., *Israel's Industrial Finances . . . A Second Look*. Washington, 1960, E.
160. ECONOMIC ADVISORY STAFF, *The Israel Economy in 1954*. Jerusalem, 1955. Mimeograph, H/E.
161. ECONOMIC PLANNING AUTHORITY, *Suggestions for Objectives and Outline of Development Plan, Four Years 1963/64–1966/67*. Jerusalem, 1962. Mimeograph, H.
162. ———, *Israel's Economic Development: Past Progress and Plan for Future*. Jerusalem, 1968, H/E.
163. FALK PROJECT FOR ECONOMIC RESEARCH IN ISRAEL, *Family Saving Survey 1957/58*, Research Paper 8, 1959, H/E.
164. ISRAEL MANAGEMENT CENTER, *The Interest Law*, Financial Management F–19, n.d., H.
165. JOINT PLANNING CENTER FOR AGRICULTURE COLONIZATION AND ECONOMIC ADVISORY STAFF, *Israel Agriculture 1953/54*. Tel–Aviv, 1955, H/E.
166. JOINT PLANNING CENTER FOR AGRICULTURE AND COLONIZATION, *Five Year Agricultural Development Plan 1966/67–1970/71*. Tel–Aviv, 1960, H.

167. Joint Planning Committee for Agricultural Colonization, *Agricultural Colonization Plan 1950–53*. Tel–Aviv, 1950. Mimeograph, H.

168. The Kaplan School of Economics and Social Sciences, *Development of Water Resources in Israel: Economic Aspects*. (Symposium in honor of Yehuda Grunfeld). Jerusalem, 1965, H.

169. Ministry of Finance, Budget Division, "The Fundamentals of Evaluating Profitability," Technical Paper No. 1, 1961. Mimeograph, H.

170. *Report of the Foreign Exchange Committee*, Jerusalem, 1953. Mimeograph, H.

171. *Report of the Public Commission on Government Value Linked Loans*, Jerusalem, 1955, H.

172. *Report of the Public Committee to Examine the Condition of Agriculture in Israel*, Jerusalem, 1960, H.

173. *Report of the Committee to Investigate the Cost of Living Allowance*, Tel–Aviv, 1966. Mimeograph, H.

The American Economic Review

THE AMERICAN ECONOMIC ASSOCIATION

Founded in 1885

• Published at George Banta Co., Inc., Menasha, Wisconsin.

• THE AMERICAN ECONOMIC REVIEW, including four quarterly numbers, the *Proceedings* of the annual meetings, the *Directory* and Supplements, is published by the American Economic Association and is sent to all members five times a year, in March, May, June, September and December.

• Membership dues of the Association are $10.00 a year, $8.00 of each year's dues being in payment of a year's subscription both to the *Review* and the *Journal of Economic Literature*. Subscriptions by nonmembers are $10.00 a year. Single copies of the *Review*, $2.00; of the *Journal*, $1.50.

• Correspondence relating to the *Paper and Proceedings*, the *Directory*, advertising, permission to quote, business matters, subscriptions, membership and changes of address may be sent to the secretary, Harold F. Williamson, 629 Noyes Street, Evanston, Illinois 60201. To be effective, notice of change of address must reach the secretary by the 1st of the month previous to the month of publication.

• Second-class postage paid at Menasha, Wisconsin. Printed in U.S.A.

Security Pricing and Investment Criteria in Competitive Markets

By JAN MOSSIN*

The theory of finance has—or should have—two basic objectives. On the one hand, the theory should be able to explain and interpret phenomena in financial markets. On the other hand, the theory should assist management in making the best decisions with respect to the company's investment and financing decisions by providing useful analytical tools within a realistic theoretical framework.

Everyone would agree that these two aspects of the theory of finance are closely related. The reason is very simple. The most general and universally acceptable formulation of company objectives is maximization of the market value of the company's equity. Through its actions management can *influence* the market value, but is clearly unable to *determine* it completely. The market value is determined by the simultaneous interplay of supply and demand in the capital markets, where other companies also participate as suppliers of securities and where various investors participate in the demand for these securities. No theory of finance can give a satisfactory explanation of security valuation or investment behavior if it fails to take into account the relationships that exist with individual investors' portfolio decisions. This means that all the investment alternatives open to the investor must be taken into account if we want to understand his evaluation of any one of them. Market values are determined by the demand by all investors, and this leads us to establish a theory of general equilibrium in capital markets. For without such a model, management is unable to foresee the effects of alternative investment and financing decisions.

The task of specifying relations constituting such an equilibrium model is not a mean one; so I shall only be able to indicate its structure in broad outline, but I shall also, by means of examples, show how the theory can be utilized to derive meaningful conclusions.

In view of the fundamental theoretical role that the analysis of capital markets should play for the study of the corporate decisions, it is remarkable to what limited extent such an analysis has been brought explicitly into existing financial literature. Even in modern and reputable introductions to the theory of finance, the market plays a highly indirect role. A number of hypotheses are advanced concerning the way in which the market evaluates and reacts (e.g., with respect to discounting for time and uncertainty), but these hypotheses are entirely *ad hoc* and quite arbitrary, since they are not derived from any fundamental assumptions describing market equilibrium.

In the celebrated contribution by Modigliani and Miller [4], the relationship between individual portolio decisions and market valuation is brought out more clearly, even though no general equilibrium model is explicitly specified. The M-M theory will be discussed in more detail below.

During the past few years there has been a fast growing interest in models of general equilibrium in markets for un-

* Norwegian School of Economics and Business Administration, visiting professor, Graduate School of Business Administration, University of California, Berkeley.

certain claims, i.e., financial markets. Some of the interest has centered on questions of efficiency (Pareto optimality) of risk allocations under various market arrangements [1]. From the point of view of the theory of finance, however, the main concern is with the description, or characterization, of the equilibrium situation itself (regardless of whether this situation represents an efficient allocation of risk). A pioneering work here is that by William Sharpe [6]. In [5], an attempt was made to sharpen and extend some of his conclusions. The purpose of this paper is to demonstrate, by means of a slightly modified version of this model, some of its implications for the analysis of the firm's optimal investment and financing decisions.

I. *A Market Model*

The model we shall consider is very much in the classical tradition. Formally, it is a model of pure exchange, and there are two types of exchange objects: bonds, issued either by firms or by the investors themselves; and company ordinary shares. The analysis takes as given the firms' investment and financing decisions. On the basis of the relationships to be derived between these decisions and market value we can infer the effects of alternative decisions. Thus, the activity carried out by company j is described by a random variable X_j, representing the gross yield of that activity. When the outcome of any such variable becomes known, it is distributed among investors on the basis of the shares and bonds they have acquired through the market exchange. A stylized example may serve to fix ideas: Each company is a farm which has, at the time of trading, effected its investments (planting of its land) as well as its financing (share sale and borrowing). Thus, the supply of various securities are given data. The trading then determines an allocation of these securities among investors. Crop values are still unknown and represented by random variables X_j. After the harvest is completed, crop values will be used partly for amortizing debt, while the remainder is distributed to shareholders.

The most important assumption we shall make is that the bond market is perfect in the sense that everyone, firms as well as investors, can lend or borrow any amount at a given and certain rate of interest. This is the same as disregarding default risk: all claims will be paid.

With this, the variables entering the model are the following:

For company j:

X_j = gross yield (with $E(X_j)=\mu_j$, $E(X_j-\mu_j)(X_k-\mu_k)=\sigma_{jk}$);
d_j = debt;
p_j = market value of shares;
$v_j=p_j+d_j$ = market value of company;
$R_j=X_j-rd_j$ = return to shares, where r is the certain rate of return on loans.[1]

For investor i:

w_i = initial wealth;
z_{ij} = fraction owned of company j;
$m_i=w_i-\sum_j z_{ij}p_j$ = net bond holdings;

$Y_i=rm_i+\sum_j z_{ij}R_j$ = final wealth.

The market clearing conditions:

$$\sum_i z_{ij} = 1 \qquad (\text{all } j) \tag{1}$$

[1] That is, we interpret r as one plus the interest rate. This is because, formally, we assume a single period horizon of finite length. This in itself is unimportant, however: the important limitation is that we consider only a single decision—or market—date, rather than a sequence of decision dates. The model and the variables are easily reinterpreted so as to represent an infinite horizon. In that case, r must represent the interest rate (per time unit) itself, while X_j must be taken as an infinite *flow* of yield (or net profit before interest payments), in principle constant through time, but unknown as of the market date.

(2) $$\sum_i m_i = \sum_j d_j$$

Apart from the assumption of a perfect bond market, everything written down so far is a definition or accounting identity. Nevertheless, we are already in a position to show a very simple result. This is the well-known Proposition I of Modigliani and Miller [4], namely, that total company value is independent of the level of debt. For by substitution of the budget equation and the definition of R_j, we can write final wealth as

(3) $$\begin{aligned} Y_i &= r\left(w_i - \sum_j z_{ij}p_j\right) \\ &\quad + \sum_j z_{ij}(X_j - rd_j) \\ &= rw_i + \sum_j z_{ij}(X_j - rv_j). \end{aligned}$$

In other words: regardless of the portfolio the investor chooses to buy, the probability distribution for final wealth will be independent of how company values are composed of share values and debt. Then, suppose a change occured in the d_j's. If $\sum_i z_{ij} = 1$ before the change, i.e., demand for shares equalled the supply of shares, the equality would hold after the change. As a result of this and the fact that $\sum_i w_i = \sum_j v_j$ identically, excess demand for bonds would also remain at zero. In other words, if debt-equity ratios are changed, and company values remain unchanged, all markets will clear; hence the equilibrium values will in fact be independent of debt levels. Given the probability distribution for yields, the amounts of debt will be entirely incidental. This particular formulation is important: if, for some reason, a dependence exists between the funds raised by borrowing and the level of operations carried out by the firm,[2] then a corresponding dependence would be observed between debt level and market value. Yet, as it stands, M-M's Proposition I follows almost as an accounting identity, given of course, the assumption of a perfect loan market. If interest rates differ, either among companies or in relation to private loans, then clearly (3) does not hold.

We have not yet specified investors' demand for shares and bonds. In principle, these are easy to write down. By choosing different sets of values of the z_{ij}, the investor will obtain different probability distributions for final wealth. If he follows certain rules of consistency (the von Neumann-Morgenstern conditions) in his choice among such probability distributions, then there exists a utility function $U_i(Y_i)$ such that the preferred values of the z_{ij} are those that maximize expected utility, $E[U_i(Y_i)]$. His demand is thus described by the relations

(4) $$\begin{aligned} \frac{\partial E[U_i(Y_i)]}{\partial z_{ij}} &= E[U_i'(Y_i)(X_j - rv_j)] \\ &= 0 \qquad (\text{all } j) \end{aligned}$$

Formally, the model is now complete. Under normal conditions we may assume that our equations determine the market solution, i.e., the market value of each company and the corresponding equilibrium portfolios.

To see the potentiality of the model for generating meaningful hypotheses, I shall proceed to specify the preference structures of investors by assuming, as an illustration, that these can be represented by quadratic utility functions:[3]

[2] As, for example, in Diamond's model [2], to be discussed later.

[3] I shall refrain from another review of the various shortcomings of, and objections to, quadratic utility functions; any student of the literature on risk taking will be thoroughly familiar (if not fed up) with the topic. The justification here is clearly that of providing a particularly simple and illuminating application of the general theory. It should be added, though, that quadratic utility functions do allow an unconstrained Pareto optimal risk allocation to be achieved by means of markets for bonds and ordinary shares only (see [1]).

$$U_i(Y_i) = Y_i - c_i Y_i^2$$

This will make it possible to derive an explicit expression for market values, and it is obvious that they will depend only on expected yields and covariances between yields.

II. *A Market Valuation Formula*

The derivation of such a formula is straightforward, although somewhat tedious. By substituting for marginal utility, taking expectations, and rearranging, we shall be able to write the demand relations on the form

$$\sum_k z_{ik}[\sigma_{jk} + (\mu_j - rv_j)(\mu_k - rv_k)] = (\mu_j - rv_j)\left(\frac{1}{2c_i} - rw_i\right) \quad (\text{all } i, j)$$

which give, for each investor, a system of linear equations with constant coefficients. Taking now the j'th equation for each investor and summing (over i), we obtain

$$\sum_k [\sigma_{jk} + (\mu_j - rv_j)(\mu_k - rv_k)] = (\mu_j - rv_j)\left(\sum_i \frac{1}{2c_i} - r\sum_k v_k\right) \tag{5}$$

On the left-hand side, the z_{ik} have dropped out because of (1). With respect to the right-hand side we have, by summing the budget equations over i:

$$\begin{aligned}\sum_i w_i &= \sum_i m_i + \sum_i \sum_k z_{ik} p_k \\ &= \sum_k d_k + \sum_k p_k \\ &= \sum v_k.\end{aligned}$$

The term $r(\mu_j - rv_j)\sum_k v_k$ then drops out on both sides in (5), and we are left with

$$(\mu_j - rv_j)\left(\sum_i \frac{1}{2c_i} - \sum_k \mu_k\right) = \sum_k \sigma_{jk},$$

which, when solved for v_j, gives

$$v_j = \frac{1}{r}\left[\mu_j - \frac{\sum_k \sigma_{jk}}{\sum_i \frac{1}{2c_i} - \sum_k \mu_k}\right]. \tag{6}$$

Here we have a surprisingly simple formula for the company's market value, expressed solely in terms of given market parameters. The interpretation of the formula is also very simple. The second term is, as we shall see, a correction term for risk; so that the market value is computed simply by taking expected yield, allowing a certain deduction for risk, and then discounting the difference at the sure rate of interest.[4] Of course, formulae of this nature have often been suggested in earlier literature, but here we have *derived* the formula from the description of market equilibrium. It is to be emphasised, however, that the formula makes no claim on generality, depending as it does on the particular preference structure we have assumed.

The correction term for risk is seen to consist of two factors. The first one is the magnitude

$$b_j = \sum_k \sigma_{jk},$$

i.e., the sum of the covariances of yield with the yields of all other companies. Thus, in the case of quadratic utilities, this is the relevant risk measure associated with company j's activities. It can be interpreted as the contribution of company j to the market's total variance, since we obviously have

$$\sum_j b_j = \sum_j \sum_k \sigma_{jk} = \operatorname{var}\left(\sum_j X_j\right).$$

It is not very surprising to learn that a company's risk cannot be measured by its own variance (σ_{jj}) alone, but also depends on its correlation with other firms. In

[4] Note that if we consider the model's time span as infinite, with r being the rate of interest (fn. 1), then $1/r$ becomes the appropriate discount factor.

traditional approaches to financial analysis, this dependence is often missed, however. Loosely speaking, a rainwear manufacturer is worth relatively more if everybody else produces ice cream and suntan lotion. We even observe that if the yield of a company is sufficiently negatively correlated with other companies' yields, then b_j may well become negative, i.e., its risk premium is negative. Since holdings in such a company strongly reduce the variances of yield on investors' portfolios, demand will be so large as to raise the market value of the company above its expected present value.

The second factor, which serves as a weighting factor for the company risk factor b_j, is

$$R = \frac{1}{\sum_i \frac{1}{2c_i} - \sum_k \mu_k}$$

This factor is seen to be the same for all companies and can be given an interpretation as market risk aversion. I shall skip the details, but it can readily be shown that R is the harmonic mean of investors' expected risk aversion, where by risk aversion we mean the Pratt-Arrow risk aversion measure

$$R_i = -\frac{U_i''}{U_i'}$$

Then,

$$R = \frac{1}{\sum_i E\left(\frac{1}{R_i}\right)}$$

It is of some interest to observe that the introduction of additional investors, even if each possesses a very small amount of wealth, may influence company values considerably. The explanation of this non-intuitive result is quite simple, however; even though an investor may have an insignificant amount of wealth, he may be able to borrow extensively and in this way affect security demand on roughly the same scale as a wealthy investor.

III. *Implications for Investment Decisions*

We are now in a position to examine some of the implications of our model for the firm's optimal investment policy. It is useful to begin by considering the M-M investment theory, and to do so we must make clear the nature of their risk class assumption, described as follows:

> We shall assume that firms can be divided into "equivalent return" classes such that the return on the shares issued by any firm in any given class is proportional to (and hence perfectly correlated with) the return on the shares issued by any other firm in the same class. This assumption implies that the various shares within the same class differ, at most, by a "scale factor." [4, p. 266]

Even though they use the expression "return on shares", it is clear from the subsequent analysis that the assumption refers to (gross) yield; not necessarily the yield on equity. The assumption therefore, is that firms can be divided into risk classes, such that two firms, j and k, belong to the same class if, and only if,

$$X_j = \alpha X_k.$$

According to M-M, this assumption "plays a strategic role on the rest of the analysis". This is because M-M set out to analyze investments undertaken by a firm in a given risk class. Now it is clear that if company j undertakes an investment which changes its yield from X_j to X_j', it will stay within the same risk class only if X_j' and X_j are proportional, i.e., if

$$X_j' = (1 + \lambda)X_j.$$

The real meaning of the risk class assumption is therefore that the M-M analysis is restricted to treat investments

which change the scale of operations only, or to use modern term, which are such that company yield changes by the same percentage in all "states of the world". The possibility of classifying different companies according to risk seems of less importance.

The theoretical basis for M-M's investment analysis is now that an investment which increases yield by a certain percentage in every state of the world must lead to the same relative increase in the company's market value:

$$v_j' = (1+\lambda)v_j.$$

Looking at the market value formula (6), we see that this is almost correct, but not entirely. It is easy to see that the new expected yield is

$$\mu_j' = (1+\lambda)\mu_j,$$

and that the same applies to all covariances (and so also to their sum), i.e.,

$$b_j' = (1+\lambda)b_j.$$

If now the risk aversion factor R remained unchanged, we would obviously have $v'_j = (1+\lambda)\ v_j$. Yet, R is not entirely unaffected by changes in yields. We see that R depends on expected aggregate yield $\sum_k \mu_k$, and this will necessarily change when the expected yield of any one company changes. In general, risk aversion will vary with yield, so that that strict proportionality will not be observed.[5] What we shall do, however, is to treat this deviation as a minor and second order effect by assuming that the number of firms is so large that the change in the risk aversion is negligible. This is essentially the same as assuming that the companies regard themselves as pricetakers in the security markets.

[5] Since quadratic utility functions exhibit increasing risk aversion we have v'_j somewhat less that $(1+\lambda)v_j$. Perhaps a more plausible hypothesis is decreasing risk aversion; in that case the inequality would be reversed.

Under this assumption it is easy to see from the first order conditions (4) that the proportionality relation holds for abritrary utility functions: if a change in one company's yield does not appreciably effect the investor's marginal utility, then equations (4) are satisfied only if v_j and X_j are changed in the same proportion.

If the proportionality relation is accepted, the rest of the analysis is very simple. To illustrate, we consider an investment priced at I which will increase yield (and so also market value) by 100λ pct., to be financed by borrowing. The new level of debt then becomes

$$d_j' = d_j + I,$$

and the market value of equity

$$\begin{aligned} p_j' &= v_j' - d_j' \\ &= (1+\lambda)v_j - d_j - I \\ &= p_j + \lambda v_j - I. \end{aligned}$$

Consequently, the investment will be acceptable only if

$$\lambda v_j \geqq I,$$

or, equivalently, if

$$\frac{\lambda\mu_j}{I} \geqq \frac{\mu_j}{v_j}$$

This is the M-M Proposition III: the criterion for undertaking the investment is that its expected rate of return is at least as high as the current expected rate of return of the company. As shown by M-M, this rule holds regardless of the method of financing the investment; borrowing was meant here as an illustration only.

The M-M analysis is certainly not trivial, but since it can handle only very special types of investments, its applicability is limited. The alternative suggested here is to buy increased generality with respect to the nature of investments at the cost of decreased generality with respect to the nature of preferences. By means of an explicit market valuation for-

mula like (6), we are able to analyze investments with completely arbitrary yield characteristics. To see how this is done, consider again loan financing of an investment I, which will generate a yield represented by the random variable Z, which has mean μ_z and covariances σ_{zk} with all other yields (including X_j).[6] With this investment, company yield will be the variate

$$X_j' = X_j + Z,$$

and it is easily verified that

$$\mu_j' = \mu_j + \mu_{zj}$$

$$\sigma_{jk}' = \begin{cases} \sigma_{jk} + \sigma_{zk} & \text{for } k \neq j \\ \sigma_{jj} + 2\sigma_{zj} + \sigma_{zz} & \text{for } k = j \end{cases}$$

and consequently

$$b_j' = b_j + b_z$$

where $b_z = \sum_k \sigma_{zk} + \sigma_{zj}$, which is the relevant risk measure for the investment when undertaken by company j. The new market value of the company will be

$$v_j' = v_j + \frac{1}{r}(\mu_z - Rb_z),$$

so that the market value of equity becomes

$$p_j' = p_j + \frac{1}{r}(\mu_z - Rb_z) - I$$

Thus the investment will be undertaken only if

$$\frac{1}{r}(\mu_z - Rb_z) \geqq I,$$

or, equivalently, if

$$\frac{\mu_z - rI}{b_z} \geqq R. \tag{7}$$

[6] We have to regard other firms' investments as given independently of the decision of company j. If this assumption is not made, we are conceptually in a very complicated oligopoly situation, but one whose effects on the descisions to be made appear as rather second order.

From currently available data on the firm's activity we can compute the value of R as

$$R = \frac{\mu_j - rv_j}{b_j},$$

and assuming, as before, that changes in R are negligible, we can write the investment criterion as

$$\frac{\mu_z - rI}{b_z} \geqq \frac{\mu_j - rv_j}{b_j}. \tag{8}$$

Again we have, on the right-hand side, objective data on the firm's current activities, its expected yield net of interest costs divided by its current risk factor. On the left-hand side we have corresponding information on the proposed investment, expected net yield divided by its own risk factor. It can easily be shown that this criterion is independent of the method of financing.

A Remark On the "Social Rate of Discount" Debate.

Without going into details, this debate can be summarized as follows. One argument is that individual enterprises are too small to be willing to undertake socially desirable risky investments. The counterargument is that with a market for shares, investors are able to diversify to the extent that the socially optimal risk level is attained. It is seen that the model presented here supports the latter argument. Assume that the investment opportunity Z is technologically independent of the existing firms in the sense that its stochastic properties are the same regardless of which company undertakes it; this means that the left-hand side of inequality (7) will have the same value for all companies. Also, R is the same for all companies, so that the criterion for undertaking the investment is the same for all companies. Even if all existing companies were merged into one, the criterion would not change.

Hence, the only barriers to a socially optimal risk level are imperfections in the stock market.

IV. *Conclusion*

I believe that the analysis so far is sufficient to demonstrate both the usefulness and the necessity of tying the theory of market adjustment in with the analysis of firm decision making. The results we have obtained are admittedly specific to the particular preference structure we have assumed, but nevertheless serve to illustrate the structure of the theory. It may also prove possible, by closer analysis of the general model, to develop operational results of the same nature.

Apart from the preference structure, there are two particular directions in which generalization and extension of the theory seem useful and even possible. One is the possibility of removing the assumption of a perfect loan market without default risk. Formally, this requires a substantial increase in the number of markets, since we shall then have to distinguish among as many different types of bonds as there are companies (and possibly even more), each with its particular price. Analytically, such a formulation poses formidable problems. Some work has been done on models incorporating default risk [7]; interestingly, it appears that at least parts of the M-M theory can be rescued also in this case.

A second major task is to make the analysis less partial by extending the model's domain of decision. The theory presented above gives no real explanation of how or why companies have arrived at the yield distribution and the amount of debt they happen to have; rather, we analyze marginal changes in company activity. Clearly, there must be some kind of discretion involved in the relationship between the liability side of the firm's balance sheet and the yield made possible by the corresponding resources on the asset side. One approach to some of the problems involved is represented by the important contribution of Peter Diamond [2]. In Diamond's model there is, in addition to markets for stocks and bonds, a market for a factor of production which firms acquire from the public. Company debt then consists exclusively of claims to compensation for factor input. Company yield becomes a random variable which depends upon the input level:

$$X_j = f_j(d_j, \theta)$$

Here θ is a random variable (the state of nature) representing the stochastic nature of production. A formulation along this line represents a step in the right direction, and it turns out that in such a model the characterization of the firm's optimal investment policy becomes a strangely intricate problem on which future research may hopefully have more to say.

References

1. K. Borch, "General Equilibrium in the Economics of Uncertainty," in K. Borch and J. Mossin, eds., *Risk and Uncertainty.* London 1968.
2. P. A. Diamond, "The Role of a Stock Market in a General Equilibrium Model with Technological Uncertainty," *Amer. Econ. Rev.*, Sept. 1967, *57*, 759–76.
3. E. M. Lerner and W. T. Carleton, *A Theory of Financial Analysis.* New York 1966.
4. F. Modigliani and M. H. Miller, "The Cost of Capital, Corporation Finance, and the Theory of Investment," *Amer. Econ. Rev.*, June 1958, *48*, 261–97.
5. J. Mossin, "Equilibrium in a Capital Asset Market," *Econometrica*, Oct. 1966, *34*, 768–83.
6. W. F. Sharpe, "Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk," *J. Finance*, Sept. 1964, *19*, 425–42.
7. J. E. Stiglitz, "A Re-Examination of the Modigliani-Miller Theorem," *Amer. Econ. Rev.*, Dec. 1969, *59*, 784–93.

Interstate Wage Differentials: A Cross Section Analysis

By GERALD W. SCULLY*

The purpose of this article is to inquire into the nature of interstate wage differentials among manufacturing production workers in 1958 with the focus of attention being on the classic case of the differentials among northern and southern workers. This choice is motivated by the relative severity and longevity of the wage differential, and its apparent importance as measured by the previous attention economists have accorded it.

One of the facts of economic life seems to be the general phenomenon of production workers in southern states receiving lower wage rates than their northern counterparts for work that is similarly performed. To be sure, the wage differential has narrowed over time. According to an early study by J. W. Block [3], the difference in pay between northern and southern workers in 1907 was about 100 percent. By 1946, the wage differential had narrowed to about 25 percent [3]. Other investigators [7], [8], [17] have confirmed the 20 to 25 percent overall magnitude of the wage differential for 1947. However, the narrowing trend in this period was not uniform. The period 1919 to 1929 evidenced relative stability of regional wage differentials [4], while the period 1929 to 1947 witnessed an improvement in the southern wage position [7]. Evidence for the 1947–54 period is conflicting. Martin Segal [17] noted some narrowing for a majority of manufacturing industries, while Victor Fuchs and Richard Perlman [7] found relative stability. Most investigators agreed that the overall wage differential in manufacturing for 1954 was in the neighborhood of 20 percent [8]. There had been little change in the differential by 1963 [5].

Much of the thrust of previous investigations has been confined to estimating the magnitude of the regional wage differential. Few attempts have been made to identify the factors which have caused and have served to perpetuate wage differentials, and rigorous testing of hypotheses has been lacking. In the more recent literature on the subject, three direct influences have been identified. Fuchs and Perlman, following Frank Hanna's methodology [12], were able to demonstrate that a portion of the wage differential was explained by regional differences in industry mix [7]. However, even after the adjustment for industry mix had been made, a sizeable differential remained unexplained. Segal suggested that interregional variations in the policies of unions and management resulted in the establishment and perpetuation of regional wage differentials [17]. Lowell Gallaway's article suggested that differing factor proportions, sustained in the long run by barriers to the free interregional flow of factors of production, was the most probable cause of the differential [8].

I. *The Hypothesis and the Data*

According to neoclassical theory, a wage differential between the southern and

* The author is assistant professor of economics at Southern Illinois University. He is indebted to Lowell E. Gallaway, Matityahu Marcus, Stanley Masters, Rajindar Koshal and the referree for suggestions. The article is based on his doctoral dissertation at Rutgers University. All disclaimers of responsibility for remaining errors apply to these individuals.

northern states should not exist given perfect competition and long-run equilibrium.[1] However, despite the expectations from neoclassical theory, wage rates have not been equalized. Conventional economic theory and previous research in the area of wage differentials and the economics of human capital and discrimination suggest three major sources which could account for the failure to achieve North-South wage uniformity and these lead to an a priori hypothesis. One major source of the variation in wage rates found among production workers in northern and southern states could be barriers to the free flow of resources among the regions, which implies different factor proportions. A second cause may be found in the non-homogeneity of the labor force, i.e., interstate variations in the amount of human educational capital manifest in workers. Finally, certain institutional factors may impede wage uniformity. Such impediments include differences in bargaining strength and labor market discrimination against minority segments (nonwhites and females) of the labor force. Therefore, variations in an industry's wage rate for production workers among the states may be related to five economic factors: (1) variations in the capital-labor ratio, (2) variations from state to state in the human educational capital manifest in production workers, (3) differences in union activity among the states, (4) variations in the percent of nonwhite production workers, and (5), variations in the percent of female production workers.

The hypothesis may be stated symbolically as

$$(1) \quad w_{ij} = f(CL_{ij}, HC_{ij}, PU_{ij}, PNW_{ij}, PF_{ij})$$

where w, CL, HC, PU, PNW, and PF denote, respectively, the average hourly wage rate paid to production workers,[2] the capital-labor ratio, the human capital-labor ratio, unionized production workers as a percent of the work force, nonwhites as a percent of the labor force, and females as a percent of the labor force. The subscripts i and j signify two-digit industry and state,[3] respectively.

Turning now to the rationale embodied in (1), it is clear that the existence of geographic wage differentials may be partly explained by different regional factor proportions, i.e., capital-labor ratios in the South may be lower than those found in northern industries. Such differential capital-labor ratios would imply lower wage rates in the South. This can be demonstrated by considering the general linear homogeneous production function

$$(2) \qquad P = F(C, L),$$

where P denotes output and C and L are capital and labor inputs, respectively. The respective marginal products of the factor inputs are obtained by differentiating (2) with respect to each factor input.

$$(3) \quad MP_L = \partial P/\partial L; \quad MP_C = \partial P/\partial C.$$

By virtue of Euler's Theorem, the factors of production are remunerated by an amount equal to their respective marginal products and the sum of factor remunerations exactly exhausts the total output.

$$(4) \quad (\partial P/\partial L)L + (\partial P/\partial C)C = P.$$

[1] See George Borts [4] for a more complete discussion.

[2] In this study, average hourly earnings, including overtime, of manufacturing production workers were employed. Use of annual earnings data would yield similar results. See Fuchs and Perlman [7, p. 292].

[3] The states which are generally recognized as being in the North-South area are: Maine, New Hampshire, Vermont, Massachusetts, Rhode Island, Connecticut, New York, New Jersey, Pennsylvania, which constitute the northern states; and Delaware, Maryland, District of Columbia, Virginia, West Virginia, North Carolina, South Carolina, Georgia, Florida, Kentucky, Tennessee, Alabama, Mississippi, Arkansas, Louisiana, Oklahoma, and Texas, which constitute the southern states.

$$w = \partial P/\partial L; \qquad r = \partial P/\partial C, \tag{5}$$

where w and r denote the wage rate and the returns to capital, respectively.

The relationship between labor productivity (and implicitly the wage rate) and the capital-labor ratio remains to be shown. Since the production function, $P = F(C,L)$, is assumed to be linear homogeneous in both factor inputs we may write

$$\begin{aligned} P &= F(C, L) \\ &= LF(C/L, 1) \\ &= Lf(C/L). \end{aligned} \tag{6}$$

A priori our expectation is that, if the differences in the wage rate are related to differences in the capital-labor ratio, there should be a positive association between the variables, i.e., $\partial w_{ij}/ \partial CL_{ij}>0$. This defines the expected sign of the CL_{ij} variable shown in expression (1).

Of course, for empirical purposes some statistical measure of the capital–labor ratio by state and two-digit industry is required. This is a relatively straight forward matter, since data by state and two-digit industry are available from the 1958 *Census of Manufactures*.[4] The specific capital stock measure which is used is gross capital.[5] Combining this with the *Census of Manufactures* man-hours of production workers data permits calculating capital-labor ratios at the appropriate level of disaggregation.

In recent years, considerable interest in the importance of human capital has developed particularly following the contributions of Theodore Schultz [15] and Gary Becker [1]. To my knowledge, no one has attempted to explain variations in interstate wage rates in terms of variations in human capital. On theoretical grounds the relationship between the wage rate and the ratio of human capital to labor is the same as that posited for the physical capital-labor ratio, namely, $\partial w_{ij}/ \partial HC_{ij}>0$. However, the statistical measurement of the HC_{ij} variable is somewhat more difficult.

Because of the importance of the method by which the human capital variable is measured, it must be discussed briefly (a more complete discussion is contained in the Appendix). A variety of factors determine the amount of human educational capital embodied in production workers: (1) human educational capital is related to age, since median school years completed and real expenditures per enrolled pupil for education have grown over time; (2) it is a function of race in view of racial differences in expenditures on education and length of schooling; (3) it is related to sex, since females in the manufacturing labor force generally have more median school years completed than do males; (4) it is a function of the state or region in which a person was educated, since length and quality of education vary from state to state.

The human capital estimates are derived from 1960 *Census of Population* data on the age, race, sex, state or region of birth and median school years completed and from deflated *Statistical Abstract* data on expenditures per enrolled pupil. Specifically, I constructed a matrix of production work-

[4] Source: "Supplementary Employee Costs, Costs of Maintenance and Repair, Insurance, Rent, Taxes, and Depreciation and Book Value of Depreciable Assets, 1957." See [19].

[5] Three measures of the capital stock can be obtained from the data; i.e., gross capital stock, which is the gross book value of plant and machinery on December 31, 1957; net capital stock, which is the gross book value on December 31, 1957 minus accumulated depreciation and depletion charged in 1957; and gross capital plus rented capital stock. A reasonable assumption is that both depreciation and depletion and rented capital stock are endogeneously related to the gross capital stock. Therefore, the gross capital stock measure was selected for calculating the capital-labor ratio. The assumption was checked with the data. The coefficients of correlation between the gross, net, and gross plus rented capital stocks were very high.

ers based on the aforementioned demographic characteristics. Each cell of the matrix contained an estimated group of production workers of a particular age, sex, race, and state or region of birth. Each cell was then assigned its appropriate summed educational expenditures which began at the time of schooling and terminated at the individual's median school years completed. In this manner an estimate was obtained of human educational capital per production worker man-hour by state and two-digit industry.

Many of the previous studies of regional wage differentials suggest that regional differences in unionization may be an important explanatory variable. The presumption is that differences in the level of union activity will be positively related to wage levels, i.e., $\partial w_{ij}/\partial PU_{ij}>0$. However, no specific test has been introduced in these inquiries to examine the hypothesis that the extent of unionization is related in a statistically significant way to the interstate wage pattern. This shortcoming is the result of the lack of data on either union membership or the extent of collective bargaining agreements by state (region) and two-digit industry. Union membership data by state and two-digit industry are nonexistent. Limited data on collective bargaining agreements for some two and three-digit industries by region (but not by state) are available in the *BLS Industry Wage Surveys*, but the data were too spotty for use here. Consequently, *BLS Industry Work Stoppages* data are used to test for the effect of unionization.[6]

Finally, we have the race and sex variable incorporated in our basic hypothesis. Economists in recent years have shown an increasing interest in the economic nature and consequences of discrimination. Our concern is to specify the theoretical possiblity of wage discrimination against minority labor[7] and to test for its existence. Consider the following simple model in which it is assumed that: the minority labor group (i.e., nonwhites or females) is employed in significant numbers in the industry; labor is homogeneous except for its identifying minority characteristic (i.e., race or sex); and factor and product markets are perfectly competitive.

A firm's proceeds, S, consist of the money value of its output P, and may be defined as

$$S = G(C, L). \tag{8}$$

Profits, π, obtained from employing variable L with fixed C, are defined as

[6] Two problems arise in using these data. First, the data are for work stoppages in states having twenty-five or more stoppages in any given year. Thus, in states where union activity was particularly low, no data are available. Of the twenty-six North-South states, data are not available for Maine, Vermont, Delaware and Mississippi. A reasonable assumption is that union activity in New Hampshire will serve as a proxy for Maine and Vermont, Maryland for Delaware, and North Carolina for Mississippi. The second problem is a bit more serious. Union activity, expressed as the percentage of workers involved in work stoppages which is the specification of the union variable used in this study, may vary from year to year for a variety of reasons. Collective bargaining agreements are often concluded for a period longer than a year, and the variable may shift over the trade cycle. Thus, economic factors in any given year may be such that the number of workers involved in work stoppages may be ususually large or small. To obviate this difficulty, the mean value of the workers involved in work stoppages over the period 1956–60 is employed. Until other data become available this variable would seem to be an adequate measure of the influence of union activity on interstate wage differentials. It should be noted that it has its weaknesses. First, it may measure more than pressure to bid up wage rates; work stoppages can result from unionization membership drives and jurisdictional disputes. Second, a higher percentage of workers involved in work stoppages may occur in low wage paying states reflecting an effort to bring wage rates into parity with other regions. While these two limitations are recognized, it is felt that higher percentages of workers involved in work stoppages will reflect the operation of trade union market forces and will be positively associated with higher wage rates. This is consistent with the positive sign which has been posited for $\partial w_{ij}/\partial PU_{ij}$.

[7] I am grateful to Lowell E. Gallaway for suggesting this approach to me.

$$\pi = G(C, L) - wL, \tag{9}$$

where w denotes the wage rate. When (9) is differentiated with respect to L, the results set equal to zero, and solved, the condition is obtained that profits are maximized when $w=\partial G/\partial L$. To incorporate minority and majority labor inputs, (9) is rewritten as

$$\pi=G(C, L+L^*)-[wL+(w-d)L^*], \tag{10}$$

where the superscript * denotes the minority segment of the labor force and the unsuperscripted variables are for the majority segment of the labor force.

The differential d is determined by the employer's "taste" for discrimination. In the case of the hiring of nonwhites, the employer incurs a cost in the sense of personal discomfort. Do employers experience any personal discomfort in hiring females? Well, I suppose that depends on whether the employer is a misogynist. However, many employers do have a "taste" for discriminating against females: because of actual or expected higher turnover rates; because they envision single females marrying or leaving the job; married females becoming pregnant and leaving; because to pay equivalent rates may lower the morale (productivity) of male labor; and because of difficulties in disciplining female workers, a factor which arises out of the traditional social relationship between men and women. Other factors than those described may cause disutility to the employer who hires a minority segment of the labor force, but without going deeper into this treacherous territory, it is suggestive of the fact that a disutility function can be specified for employers who hire minority labor. Therefore,

$$D = \phi(L^*), \tag{11}$$

where D signifies the disutility incurred by hiring minority labor. To incorporate the employer's "taste" for discrimination, (10) is rewritten as (12) on the assumption that D can be translated into a money equivalent.

$$\begin{aligned}\pi = G(C, L + L^*) - \phi(L^*) \\ - [wL + (w - d)L^*].\end{aligned} \tag{12}$$

To maximize profits the employer must maintain a differential equal to his marginal taste for discrimination.

$$d = \phi'(L^*), \tag{13}$$

yielding a minority wage rate of

$$w^* = (w - d). \tag{14}$$

Condition (14) implies that the majority segment of the labor force will receive a wage rate which will equal the combined marginal value product of both segments of the labor force (in view of the homogeneity assumption),

$$w = \partial G/\partial(L + L^*), \tag{21}$$

while the minority segment of the labor force will receive a wage rate less than the marginal value product of labor.[8]

Data on nonwhites and females as a percent of the labor force by two-digit industry and state are available in the 1960 *Census of Population*. These are the variables used in this analysis. The a priori expectation is that for both segments of the minority labor force the sign of the regression coefficient will be negative. Some questions arise concerning the interpretation and empirical validity of the variables, particularly the race variable. The function of the variables in the equation is to test for the existence of a wage differ-

[8] In view of these results, clearly, π will be higher the larger is the value of d. With profits higher in the industry that discriminates there would be a tendency, *ceteris paribus*, for resources to be reallocated toward that industry. However, on the assumption that D can be translated into a money equivalent, these extra profits accruing to those who hire minority labor are treated as costs that must be recovered if minority labor is to be employed. Assuming that disutility is constant among employers, there would be no resource reallocation toward industries who hire minority labor.

ential unfavorable to these two minority segments of the labor force. A direct test for wage discrimination is not possible, since data on wage rates for nonwhites and females is not available. The dependent wage variable which is utilized is average hourly earnings. If a wage differential exists, it will produce variations in the average wage rates. States employing higher percentages of nonwhites and/or females will tend to have lower average wage rates in view of the higher weight given to the hypothesized lower nonwhite and/or female wage rates. The following questions of empirical validity are of concern. First, is it reasonable to assume labor homogeneity except for race? Second, the race variable may be collinear with the human capital variable (making this question related to the first). Third, discrimination may differ from state to state. If low wage rates are received by production workers in industries that employ large percentages of nonwhites, the low wage rates may be a function of low levels of education (inferior nonwhite education) or discrimination or some combination of the two. The physical and human capital variables should account for most of the variation in the wage rate, if discrimination is of no importance. However, if the nonwhite variable significantly accounts for some of the variation beyond that explained by nonhomogeneous labor (i.e., labor which is associated with differing levels of physical and human capital which affect their respective marginal product schedules), some justification exists for concluding that wage discrimination is present. Regarding the second question; if the race and human capital variables are collinear, that in itself is an interesting result which should be noted. Finally, the possibility of different degrees of discrimination among the states cannot be rejected out of hand. However, to take cognizance of this possibility would involve estimating the parameter $\phi'(L^*)$ which in itself entails substantial empirical difficulties. Therefore, we choose to run the risks involved in employing the race variable as already defined.

II: *Empirical Results*

The previous discussion provides a basis for testing empirically the hypothesis embodied in expression (1). Specifically, (1) suggests a least squares regression equation of the form:

$$w_{ij} = a + b(C/L)_{ij} + c(HC/L)_{ij} + d(U/L)_{ij} - e(NW/L)_{ij} - f(F/L)_{ij} + u.$$

Where

w = the average hourly wage rate paid to production workers calculated from 1958 *Census of Manufactures* data.

C/L = the capital-labor ratio calculated from 1958 *Census of Manufactures* data.

HC/L = the human educational capital-labor ratio calculated from 1960 *Census of Population* and 1958 *Census of Manufactures* data.

U/L = the percent of the labor force (average over 1956–60) in the industry involved in work stoppages calculated from *BLS Analysis of Work Stoppages* data.

NW/L = nonwhites as a percent of the industry labor force calculated from 1960 *Census of Population* data.

F/L = females as a percent of the labor force calculated from 1960 *Census of Population* data.

u = a random error term.

In addition, in certain of the regression equations various dummy variables de-

signed to reflect certain regional and industry characteristics were introduced. These are defined as follows:

R = a dummy North-South variable utilized to determine if the independent variables accounted for all of the differential between regions. Dummy = 1, if state is in the North; dummy = 0, if state is in the South.[9]

$I_1, I_2, \cdots, I_n$ = industry dummy variables used in aggregated industry equations (n-1 dummies for n industries). Dummy = 1, if industry i; dummy = 0, if not industry i.

Table 1 contains the results of estimating the regressions shown in (1) for each of the seventeen individual industries. For purposes of discussion, the industries are grouped as low wage, medium wage, and high wage. In the low wage group of industries, the equations are all significant at the five percent level, at least, while in the medium and high wage industries, four of the six, and three of the six equations, respectively, are significant. The signs of C/L, HC/L, and U/L are consistent, on the whole, while the signs of NW/L and F/L are not. The most frequently significant coefficients are for HC/L and U/L. Given the heterogeneous nature of these industries, it was not expected that the same variables would be significant in all of them. Therefore, insignificant variables are dropped from the equations. However, multicollinearity among the independent variables may exist, and the presence of such a condition could cause the omitting of relevant variables, in addition to making the estimates of the coefficients inefficient. The specific test for multicollinearity used was that proposed by Donald Farrar and Robert Glauber [6]. For reasons of conserving space, these results are not presented here.[10] The equations in Table 2 are the final results after the test for multicollinearity. Variables that were excluded were truly insignificant, i.e., independent and insignificant. Two new variables appear in these equations. $(C+HC)/L$ is a generalized capital-labor ratio, the sum of physical and human capital. $(NW+F)/L$ is minority labor, the sum of nonwhites and females as a percent of the labor force.[11] On the whole, multicollinearity was

[9] The dummy North-South variable does not appear in the reported equations. The zero order correlation coefficient between R and the wage variable in general was very significant. However, the dummy was highly correlated with the other independent variables, and its introduction into the regression equations played havoc with the regression coefficients. Additionally, the dummy changed the $\bar{R}^2$'s of all the equations marginally. Therefore, we can conclude that the variables specified account for all of the North-South differential.

[10] An encouraging result of the multicollinearity test was that no pattern of collinearity between the various independent variables emerged. More importantly, the expectation of linear interdependence between NW/L and the other independent variables was not confirmed. However, NW/L and F/L were sometimes linear interdependent and this result was not unexpected. Given the social relationship between nonwhite males and white females, i.e., the reluctance to employ large numbers of them together, we would expect some collinearity.

[11] The standard approach to solving the multicollinearity problem is either the introduction of new information, reformulation of the variables, or rationalization of the multicollinearity. In the latter case, if the linear interdependence can be shown to be stable, then one may rationalize the efficiency of the regression equation. The first approach, that of introducing new information was not tried. Rather, the less costly approach of reformulating the variables was undertaken. Two reformulations are possible. First, physical capital and human capital can be added together so that we can speak of a generalized capital-labor ratio. This formulation is economically meaningful. How effective this transformation will be depends, on the one hand, upon the strength of association between the two variables and, on the other hand, upon the magnitude of the physical capital-labor ratio. If the C/L ratio is particularly large, the possible "corrective" effects of the addition of HC/L may be largely washed out. The second possible reformulation is the addition of NW/L to F/L. In such a case we can then refer to the effect on the wage rate of a percentage change in minority labor employed, thus treating females and nonwhites

Table 1—Individual Industry Wage Equations[1]

Industry Group	SIC Code[2]	Constant Term	C/L	HC/L	U/L	NW/L	F/L	N	$\bar{R}^2$
Low Wage Industries	22	1.1168	.0251	$.5359^{***}$	2.0977^{***}	.5028	$-.2214$	17	$.9372^{**}$
	23	2.2353	$-.0082$	$.2570^{++}$	1.7336^{++}	$-.1563$	-1.4233^{***}	16	$.9671^{**}$
	24	.9844	.0302	$.3277^{++}$	1.4234	$-.3011^{**}$	2.5481^{++}	14	$.9499^{**}$
	25	1.0939	.0410	$.5950^{**}$	1.1323	$-.4890$	.0720	15	$.8398^{**}$
	31	.9174	.0653	$.3870^{++}$	5.0617	2.2542	.0659	12	$.8104^{++}$
Medium Wage Industries	20	.8129	$.1591^{**}$	$.2203^{+}$	3.7010^{***}	$-.2252$	$-.5074$	22	$.8768^{**}$
	26	2.1996	.0086	$.3416^{*}$	$-.6444$	.0268	-1.8944^{**}	19	$.6034^{++}$
	30	1.7274	.2283	.3513	.4193	-2.0379	-1.6769	8	.9049
	32	1.1322	$.0482^{*}$	$.5725^{**}$	1.4025	$-.3571$	1.0641	15	$.8498^{**}$
	34	1.5544	.0198	$.6012^{***}$	2.2847^{***}	.7181	-1.1636	19	$.6892^{**}$
	36	1.9151	.0284	$.3855^{++}$	1.0735	-3.8495	$-.5678$	15	.4843
High Wage Industries	27	1.8118	$-.0156$	$.7349^{*}$	-1.0145	$.9123^{+}$	$-.1798$	10	.8688
	28	2.3811	$.0193^{**}$	$-.0255$	2.3327^{++}	-2.7751^{***}	$-.7030$	18	$.8853^{**}$
	29	2.2667	$.0211^{+}$	.1654	-1.0314	.3929	$-.5580$	10	$.8938^{++}$
	33	2.0848	$.0458^{**}$	.1385	$.8303^{**}$	-1.3553^{**}	$-.3482$	17	$.8318^{**}$
	35	2.0821	$-.0060$	$.4752^{+}$	.7036	.0448	-1.5931	16	.3315
	37	2.1334	$-.0034$	.2836	.4058	.0017	.7643	13	.3093

[1] Levels of significance: $^{***}=.5$, $^{**}=1.0$, $^{*}=2.5$, $^{++}=5.0$, and $^{+}=10.0$ percent levels, respectively.
[2] The industries corresponding to the Standard Industrial Classification Code are:

20 = Food and Kindred Products
22 = Textile Mill Products
23 = Apparels
24 = Lumber and Wood
25 = Furniture and Fixtures
26 = Paper and Allied
27 = Printing and Publishing
28 = Chemicals
29 = Petroleum and Coal
30 = Rubber and Plastics
32 = Leather
33 = Primary Metals
34 = Fabricated Metals
35 = Nonelectrical Machinery
36 = Electrical Machinery
37 = Transportation Equipment

Table 2—Regression Results for Final Individual Industry Equations[1]

SIC Code	Constant Term	Independent Variables			N	$\bar{R}^2$
22	1.0093	$+.0319^{+}C/L$	$+\ .5478^{***}HC/L$	$+2.0384^{***}U/L$	17	$.9203^{**}$
23	1.9873	$+.3846^{***}HC/L$	$+1.5135^{++}U/L$	$-1.1688^{***}(NW+F)/L$	16	$.9413^{**}$
24	.9928	$+.6417^{***}HC/L$	$-\ .2432^{*}(NW+F)/L$		14	$.9122^{**}$
25	1.0683	$+.6633^{***}HC/L$	$+1.2269^{+}U/L$		15	$.8245^{**}$
31	.9599	$+.3992^{*}HC/L$	$+5.9461^{+}U/L$	$+2.7615^{**}NW/L$	12	$.7881^{**}$
20	.6024	$+.1748^{***}C/L$	$+\ .1868^{+}HC/L$	$+4.1134^{***}U/L$	22	$.8587^{**}$
26	2.3545	$+.2901^{*}HC/L$	$-2.2260^{***}F/L$		19	$.5761^{**}$
30	1.8610	$+.2678U/L$	$+\ .1985^{++}(C+HC)/L$	$-1.4672^{++}(NW+F)/L$	8	$.8951^{++}$
32	1.0772	$+.0380^{++}C/L$	$+\ .7824^{***}HC/L$	$+2.0762^{***}U/L$	15	$.7985^{**}$
34	1.7295	$+.5785^{***}HC/L$	$+2.2395^{***}U/L$	$-1.4291^{+}F/L$	19	$.6689^{**}$
36	1.7750	$+.3995^{+}HC/L$	$+1.3547^{+}U/L$	$-4.0970^{+}NW/L$	15	.4843
27	1.8057	$+.6148^{***}HC/L$	$+\ .8906^{***}NW/L$		10	$.8488^{**}$
28	2.5245	$+.0137^{++}C/L$	$+\ .2665^{++}HC/L$	$+1.6395U/L$	18	$.8488^{**}$
29	2.2148	$+.0249^{***}C/L$	$-\ .7788^{++}U/L$		10	$.8848^{**}$
33	2.1788	$+.0409^{***}C/L$	$+\ .1924^{+}HC/L$	$+\ .7336^{***}U/L$ $-1.2788^{**}NW+F/L$	17	$.8252^{**}$

[1] See Table 1 for levels of significance.

not severe, i.e., the $\bar{R}^2$'s on the diagonal and their concomitant F-values were fairly low.

Table 2 shows that the coefficient of determination is significant in fourteen out of the seventeen industry regression equations. In thirteen of the fourteen significant cases, the $\bar{R}^2$ is significant at the 1 percent level, and the $\bar{R}^2$ levels are generally high. The indication, in general, is that the hypothesis embodied in expression (1) explains a very considerable portion of interstate variations in industry wage rates. The hypothesis does not explain a significant portion of the variation in the wage rate in three industries, SIC 35, 36, and 37.[12]

The regression results for the two-digit industry equations permit the following conclusions to be made about interstate wage variations. (1) The most important explanatory variable is the human capital-labor ratio. HC/L is significant in thirteen of the seventeen industries and is most consistently significant in the low wage industries. (2) Unionization differences constitute an important explanatory force in the low and medium wage industries, but do not appear to be important in the high wage industries. The reason for this result may be that among the high wage industries in general, collective bargaining agreements are negotiated nationally, thus, with exceptions of course, diminishing the force of local (state) work stoppages on the wage rate. (3) The capital-labor ratio does not emerge, at least at the two-digit level, as an important source of interstate wage variation in the low or medium wage industries, but it performs moderately well in the high wage industries. (4) The results for the nonwhite and female variables are mixed and clouded by the multicollinearity problem. However, one or more of the elements of the minority labor force was significant in ten out of the seventeen cases.

Further useful information can be obtained by pooling some of the observations to see if the hypothesis is strengthened by aggregation. A possible reason for insignificance of some of the coefficients in the separate industry equations may be the small size of the samples. Table 3 contains the equations for the aggregated industry results. Results for five groupings of the data are presented. Dummy industry variables are introduced in view of the differing intercept values of the separate industry equations.

For the low, medium, and high wage grouping of industries,[13] all of the independent variables are significant, most at the 0.5 percent level. The multicollinearity test, confirmed by the high t-values of the regression coefficients, indicates that the low and medium wage industries are free of multicollinearity. However, in the high wage industry equation, collinearity is severe between F/L and U/L and, there-

together as a minority segment of the labor force. In this case, as in the above transformation, the effectiveness of the transformation in resolving the linear interdependence problem is constrained by the same factors.

[12] The fact that the hypothesis does not yield meaningful results in these three industries does not necessarily negate the relevance of the variables in explaining the three industries' interstate wage pattern. The failure of the hypothesis to yield significant results may be caused by the level of aggregation in the industries. For example, consider the case of transportation equipment. This industry is composed of four very distinct three-digit industries: motor vehicles and equipment, aircraft and parts, ships and boats, and railroad equipment. In turn, each of these three-digit industries is composed of fairly heterogeneous four-digit industries. The values of all of the independent variables probably differ from one three-digit industry to another to a greater extent in these three industries than they do in the other two-digit industries. To obtain possibly meaningful results for these industries, therefore, the data would have to be disaggregated at least to the three-digit level. This approach was not possible given the present limitations of the data.

[13] The national wage rate for each of these industries was computed and ranked to determine low, medium, and high wage industries. Mean wages rates calculated with the sample of North-South states yielded similar rankings.

TABLE 3—REGRESSION COEFFICIENTS FOR AGGREGATED INDUSTRY EQUATIONS[1]

	Low Wage	Medium Wage	High Wage	Nondurable	Durable
Constant Term	1.2342	1.9012	2.0656	1.8342	2.0457
C/L	.0325++	.0351***	.0317***	.0294***	.0363***
NW/L	−.2506**	−.5854***		−.2640+	−.6095***
HC/L	.5658***	.4178***	.2734***	.4033***	.4167***
U/L	1.3683***	1.4221***	.2860+	.4019+	.5916***
F/L	−.4577**	−1.0762***		−1.2611***	−.2733
$(NW+F)/L$			−.4478++		
I_1	−.0131	−.3851***	.0362	−.2552***	−.8681***
I_2	−.0251	−.1332+	−.4281***	−.1095++	−.7367***
I_3	−.2558***	−.0797	−.4128***	.1721++	−.4486***
I_4	−.0547	−.2638***	−.0809	.0224	−.1118+
I_5		−.1071	−.2639***	.4995***	−.3104***
I_6				.0186	−.2706***
I_7				.0275	−.3586***
I_8				.1973*	
N	74	98	84	132	124
$\bar{R}^2$	.8824**	.6874**	.6239**	.9063**	.8922**

[1] See Table 1 for levels of significance.

fore, $(NW+F)/L$, which is not collinear with U/L, is substituted into the equation.

The range of the regression coefficients is also of interest. The regression coefficients for C/L for the aggregated low, medium, and high wage industries are .0325, .0351, .0317, respectively. These results point to a remarkable uniformity in the relationship between variations in the wage rate and variations in the capital-labor ratio. By contrast, the coefficients for the human capital-labor ratio vary. Generally, the values of the coefficients are higher in the separate low wage industry equations than in the separate industry equations for the other two industry groups, and higher in the medium wage industry equations than they are in the individual high wage industry equations. For the aggregated low, medium, and high wage industries, the values of the coefficients are. 5658, .4178, and .2734, respectively, which is consistent with the results of the separate industry equations and indicates that increases in human educational capital have a greater impact in raising wage rates in the low wage industries than in the other industries. The range for the coefficients of U/L among the individual two-digit industries is from .7336 to 5.9461. The aggregated low and medium wage industries have coefficients of similar value, and these values are much higher than the value of the coefficient in the aggregated high wage industries. Variations in work stoppages, therefore, appear to have less of an impact on wage rates in the high wage industries than they do in the low or medium wage industries. The range in the values of the coefficients of nonwhites and females as a percent of the labor force indicate wide variation in the impact of NW/L and F/L on the wage rate among the various industries. In the case of NW/L, changes in the percentage of nonwhites employed affect the wage rate more adversely in the medium wage industries than in the low wage industries. A 1 percentage point increase in the employment of nonwhites is associated with a 5.9 mill decrease in the wage rate for the medium wage industries, but with only a 2.5 mill decrease in the low wage industries. A one percentage point increase in the employment of females is associated with a 1.1 cent decrease in the wage rate

in the medium wage industries, but only with a 4.6 mill decrease in the low wage industries. Therefore, changes in the percentages of females employed more adversely affect the wage rate in the medium wage industries. The regression coefficient for the aggregated high wage industries is for the minority labor variable and the results are not strictly comparable with NW/L and F/L. For the high wage industries a 1 percentage point increase in the employment of minority labor is associated with a 4.5 mill decrease in the wage rate. Therefore, pooling the data strengthened the hypothesis considerably. The human capital variable remains the most significant explanation of interstate wage variations, although the other variables are certainly significant. Similar results are obtained in the more highly aggregated durable-nondurable industry groupings.

Finally, and in way of summary, the results of aggregating all of the two-digit industries may be considered. The pooling process increases the number of observations to $N=256$ and the result is

$$\begin{aligned} w_{ij} = {} & 2.1045 + .0321^{***}C/L - .4867^{***}NW/L \\ & + .4170^{***}HC/L + .5970^{***}U/L - .9009^{***}F/L \\ & - .6012^{***}I_1 - .5517^{***}I_2 - .3886^{**}I_3 - .9289^{***}I_4 \\ & - .6986^{***}I_5 - .3316^{***}I_6 + .1202^{+}I_7 - .3502^{***}I_8 \\ & - .3742^{***}I_9 - .2069^{**}I_{10} - .4724^{***}I_{11} \\ & - .4180^{***}I_{12} - .0977^{+}I_{13} - .2688^{***}I_{14} \\ & - .2324^{***}I_{15} - .1984^{**}I_{16}, \bar{R}^2 = .8961^{**}. \end{aligned}$$

The five independent variables are all significant at the 0.5 percent level, all with the predicted signs, and all of the dummies are significant, mostly at the 0.5 percent level. The equation explained (significant at the 1 percent level) nearly 90 percent of the variation in the wage rate.[14] The magnitude of the regression coefficients indicate that a 10 cent change in the capital–labor ratio is associated with a 3.2 mill change in the wage rate, a 1 cent increase in HC/L with a 4.2 mill increase, a 1 percentage point increase in U/L with a 6.0 mill increase, a 1 percentage point increase in NW/L with a 4.9 mill decrease, and a 1 percentage point increase in F/L with a 9.0 mill decrease in the wage rate. The relative importance of the variables judged by their respective t-values supports the results obtained with the separate industry equations and with the other aggregated industry equations. HC/L has the highest t-value, followed by C/L, U/L, F/L, and finally, NW/L.

III: *Conclusions*

Quite clearly, the empirical results which have been reported constitute a substantial confirmation of the basic hypotheses, which have been advanced as explanations of interstate wage variations in the manufacturing sector. At the two-digit industry level, the hypotheses are reasonably consistent with our expectations, but even more conclusive results are obtained with the aggregated industry equations. All five of the independent variables are highly significant in a statistical sense with the results for the capital-labor ratio, nonwhite and female variables being more conclusive in the aggregated industry equations than they were in the separate industry results.

While the results of the regression equations indicate that most of the state to state variation in production worker wage rates has been explained, we have neither examined the magnitude of the North-South differential nor have we demonstrated the portion of the wage differential associated with each of the specified factors. This can be accomplished by partitioning the all industry data into northern and southern groupings and running two new regression equations.

[14] Without the industry dummies in the equation, the results still remain conclusive.

$$\begin{aligned} w_{ij} = {} & 1.6780 + .0345^{***}C/L - 1.1957^{***}NW/L \\ & + .5229^{***}HC/L + 1.3633^{***}U/L - .9914^{***}F/L, \end{aligned}$$
$\bar{R}^2 = .7031^{**}$.

$$
\begin{aligned}
(w_{ij})_{\text{North}} = {} & 2.0240 + .0293^{***}C/L + 1.4657^{***}NW/L \\
& + .3188^{***}HC/L + .7591^{***}U/L - .7602^{***}F/L \\
& - .4670^{***}I_1 - .4647^{***}I_2 - .4172^{**}I_3 \\
& - .7631^{***}I_4 - .5609^{***}I_5 - .2704^{***}I_6 \\
& + .1783^{*}I_7 - .2187^{*}I_8 - .3311^{++}I_9 - .3005^{***}I_{10} \\
& - .4236^{***}I_{11} - .1838^{*}I_{12} - .1260^{++}I_{13} \\
& - .2076^{***}I_{14} - .1546^{*}I_{15} - .1749^{++}I_{16} \\
& \bar{R}^2 = .9425^{**}, N = 92.
\end{aligned}
$$

$$
\begin{aligned}
(w_{ij})_{\text{South}} = {} & 2.1793 + .0327^{***}C/L - .4988^{***}NW/L \\
& + .3852^{***}HC/L + .4826^{***}U/L - .7755^{**}F/L \\
& - .7115^{***}I_1 - .6846^{***}I_2 - .5640^{**}I_3 \\
& - 1.0202^{***}I_4 - .7913^{***}I_5 - .3778^{***}I_6 \\
& + .1244I_7 - .4314^{***}I_8 - .3624^{**}I_9 \\
& - .1521I_{10} - .6030^{***}I_{11} - .5355^{***}I_{12} \\
& - .1079I_{13} - .3329^{***}I_{14} - .2518^{***}I_{15} \\
& - .2677^{*}I_{16}. \ \bar{R}^2 = .8947^{**}, N = 164.
\end{aligned}
$$

All of the regression coefficients for the five independent variables are highly significant in both equations. The only sign inconsistency is that of NW/L in the northern equation, but this result is not surprising in view of the fact that nonwhites are congregated in the high wage paying states in the North. The coefficients for C/L and F/L are quite similar in both regions. In fact, there is a slightly larger impact of changes in C/L with respect to changes in w in the South. The pattern of wage discrimination against females is identical in both regions. The coefficient of HC/L is higher in the South, suggesting higher relative returns to human capital in the South. The higher coefficient for U/L in the North is susceptible to multiple interpretation. On the one hand, one might expect a declining marginal impact of unionization of wage rates, implying a higher coefficient in the South. On the other hand, a critical overall regional level of unionization may have to be reached before union activity has much of an impact on wage rates, and this suggests a lower coefficient in the South. The most revealing difference in the coefficients is that of NW/L. The sign of the coefficient in the North indicates that discrimination in the form of a white-nonwhite wage differential apparently is nonexistent. On the other hand, wage discrimination against nonwhites is present in the South. Based on other research [11], we may conclude that discrimination in the South takes the form of an unfavorable nonwhite wage differential, while in the North it takes the form of nonwhite job exclusion.

The sources of the North-South wage differential can be identified by inserting the mean values of the variables into the regression equations. Table 4 contains the results. This approach permits comparison

TABLE 4—SOURCES OF THE NORTH-SOUTH WAGE DIFFERENTIAL

Variable	North	South	Differential
Predicted wage rate, $\hat{w}$	2.1506	1.9656	0.1850
Constant Term, a[1]	2.0240	2.1793	−0.1553
$\hat{w}-a$	0.1266	−0.2137	0.3403
Regression Coefficient· Mean value			
$b\cdot C/L$.	0.1517	0.2299	−0.0782
$c\cdot HC/L$	0.3304	0.2303	0.1001
$d\cdot U/L$.	0.0611	0.0319	0.0292
$e\cdot NW/L$	0.0561	−0.0641	0.1202
$f\cdot F/L$	−0.2054	−0.1739	−0.0315
$\sum_{i=1}^{16} \bar{D}_i$	−0.2673	−0.4678	0.2008

[1] Note: the value of the constant term is determined by the reference dummy which in this case was a high wage industry, SIC 37.

of the effect of each of the variables on the wage differential. Ignoring the constant term differential, which is unfavorable to the North, the following conclusions can be drawn. (1) The net effect of the capital-labor ratio, *ceteris paribus*, is to create a wage differential unfavorable to the North. This result may appear somewhat surprising in view of Gallaway's conclusion.[15] However, this result is really not surprising. A careful examination of North-South capital-labor ratios by two-digit industries generally reveals a pattern of similar capital-labor ratios.[16] In fact, the South has a higher mean capital-labor ratio than the North. Factor proportions evidently have adjusted in the right direction; in fact, there may have been an over-adjustment. (2) The net effect of human capital is to create a wage differential unfavorable to the South. While the rate of return on human capital is higher in the South, reflecting the relative scarcity of the factor in the South, the level of human capital is nearly twice as high in the North, the net result of which is to produce a 10 cent difference in North-South wage rates. (3) Two of the variables have a minor impact on the wage differential. The pattern of discrimination against females is regionally uniform, but the slightly higher utilization of female labor in the North creates a wage differential unfavorable to the North. Trade union activity while favorable to the North, produces only about a 3 cent difference in wage rates. (4) The results obtained with the nonwhite variable indicates that the South is paying a rather high price, a 12 cent differential in wage rates, for the policy of wage discrimination. Whether the wage discrimination is justified on the basis of white-nonwhite productivity differentials is a matter for further study. The capital- and human capital-labor ratios should have accounted for productivity differentials, but there may be a variety of pre-market nonhomogeneities manifest in the nonwhite labor force which may produce productivity differentials. Whether the wage discrimination is economically justified or not, it is apparently present, and is a significant cause of the North-South wage differential. (5) Finally, the impact of the industry dummies is quite large and unfavorable to the South. One is tempted to conclude that the dummies measure the impact of the two-digit industry mix. But like the constant term in the equation, one cannot be too sure precisely what is being measured. The dummies in part may be picking up such influences as intraregional variations in technology or economies of scale, neither of which follow from theory, or in the vintage of the capital stock. However, it is believed that to a large degree the dummies measure the impact of industry mix.

Perhaps the most interesting aspect of the results is the implication that they have for the prospects of obtaining regional wage uniformity. Fairly clearly, such uniformity will depend rather heavily upon (1) the extent to which the North-South differences in human capital are eliminated and (2) the elimination of the white-nonwhite wage differential. Un-

[15] Estimating capital stock by region based on the Creamer, Dobrovolsky and Borenstein estimates, Gallaway calculated capital-labor ratios for the 20 two-digit manufacturing industries. He then examined the pattern of change in the capital-labor ratio among several broad Census regions for the period 1954–1957 and concluded that there was a general convergence in capital-labor ratios in the period. He argued, therefore, that conditions were conducive for a narrowing of the wage differential. An objection to this conclusion is simply that, while a theoretical relationship between the capital-labor ratio and the wage rate was shown, no empirical relationship between the variables was tested. Therefore, his conclusion that a narrowing of the capital labor-ratio differential was the necessary condition for a narrowing of the wage differential is somewhat speculative.

[16] For a complete discussion of differences in North-South capital-labor, ratios see [16]. Also contained therein is the complete set of data used in the regression equations.

fortunately, these are not events which one would expect to occur in the course of the normal operation of the market place. For example, ordinarily it would be anticipated that interregional factor flows would tend to equalize regional factor endowments. However, in the case of human capital, problems arise due to this factor's being embodied in the labor input to the productive process. Thus, for human capital to flow from the North to the South, labor must flow, but the combined average reward to the two factors does not encourage this. Admittedly, it would if workers with relatively large endowments of human capital were rewarded appropriately. However, in the standardized and integrated production processes which characterize manufacturing industries, the return to the human capital possessed by production workers is undoubtedly a function of the average level of endowment, and thus does not vary greatly from individual to individual.[17] Consequently, there is little inducement for human capital to flow from North to South.

Similarly, the possibility of a flow of labor with relatively low endowments of human capital from South to North producing an equalization of factor endowments is probably quite small. For labor which possesses limited quantities of human capital will find difficulty competing successfully in an industrial process which is geared to the presence of relatively high levels of this factor. In effect, southern labor with poorer education finds it difficult to compete against northern labor, and is thus discouraged from moving North. Consequently, there would appear to be little prospect of interregional flows of the stock of human capital producing an equalization of factor endowments.

With respect to the other major source of the North-South differential, the white-nonwhite wage gap, it might be anticipated that population flows would operate to lessen the impact of this factor as nonwhites would be expected to leave the South and move to the North. However, recent findings [10], [14] suggest that this type of mobility has limited potential for lessening the North-South wage differential. Specifically, there is evidence to the effect that the extent of movement of nonwhites from the South to the North is undoubtedly on the decline which means that this factor cannot be counted upon too heavily to eliminate the North-South wage differential.[18]

If normal market processes are not likely to operate to produce regional wage uniformity, what will? It would seem from our evidence that the factors which would be most likely to produce such uniformity are political and social in character rather than economic. Specifically, it would seem that political and social decisions designed to increase the stock of southern human capital through additional expenditures on public school education would be useful. To some extent this has already been accomplished through federal shifting of

[17] What happens here is that although the marginal reward to human capital is greater in the South, the average return to each body possessing this factor is lower than in the North. Since wage levels tend to be standardized in manufacturing type pursuits, employers are not in a position to reward differential endowments of human capital above the average. Nor, in all probability, should they, for it would seem likely that productivity levels are little affected by the presence of a single "over-endowed" individual.

[18] The question naturally arises as to why the flow of nonwhites from South to North should decline. Various evidence indicates that the reverse flow from North to South among nonwhites is quite substantial. Consequently, as the base of the nonwhite population increases in the North, the net flow from South to North tends to decline. For example, Gallaway [10] shows that between 1957 and 1960, 5.4 percent of Negro males migrated from South to North compared to 6.7 percent of white males. Similarly, 3.8 percent of Negro males migrated in the other direction, compared to 2.3 percent of white males. If such migration rates continue, it is clear that little can be expected from population flows as a factor in reducing the North-South wage differential.

funds for public education. However, this is a process which will require at least a generation to have its full effect since little can be done in this fashion to eliminate the differential human-capital endowments which now exist among those who have already passed through the education process.

As to the white-nonwhite wage gap, this is perhaps the most intractable problem facing our society today and it clearly involves social and political decisions of the first magnitude if it is to be eliminated. Thus, our results can probably best be summed up in two propositions:

The North-South wage differential as it now exists is primarily noneconomic in character; arising out of a social and political milieu which has generated differential endowments of human capital and a white-nonwhite wage differential which is not particularly amenable to elimination by the normal operation of the labor market.

This wage differential is likely to continue until basic decisions are made in the political and social arena which will eliminate these two factors in our society.

Appendix

Methodology of Constructing Human Capital Estimates

The general view is that human capital takes two forms: first, the education embodied in labor; second, the value of work experience. For the purpose of this study only the former was considered. While an estimate of human capital in the form of work experience would be useful, it is believed that its determination on an interstate and interindustry basis would be difficult, if not impossible, in view of the paucity of data. On the other hand, data are available for estimating human capital in the form of the value of education embodied in the labor force.

The 1960 *Census of Population* (hereafter, *Population*) contains data on the age distribution of the labor force by state (Table 128). The definition of the labor force differs from the *Census of Manufactures* (hereafter, *Manufactures*) definition of production workers and appears to be more akin to the "all employees" universe. However, use of the all employees data should not seriously bias the estimated age distribution of production workers. Estimates of the age distribution by industry i and state j were obtained from the two following expressions, where LCP denotes the *Population* labor force, LCM the actual number of *Manufactures* production workers, a, the age class interval, where $a=1, \ldots, 9$ (i.e., there are nine age class intervals ranging from 14–17 to 65 and over), and $\hat{L}$ the estimated number of *Manufactures* production workers in each age class. Therefore, we have

$$e_a = LCP_a \Big/ \sum_{a=1}^{9} LCP \tag{1}$$

$$\hat{L}_a = e_a \sum_{a=1}^{9} LCM \tag{2}$$

Table 122 of *Population* contains data on white and nonwhite employment by industry and sex. Nonwhites are calculated as a percentage of total employment in industry i and state j, and then the percentage is multipled by LCM to obtain an estimate of nonwhite production workers. The nonwhite production workers are distributed by age class intervals as the total population is distributed with the result that four race-sex estimates of LCM_{ij} are generated.

To determine the median school years completed, and the amount of school expenditures manifest in the production workers of each race-sex-age cohort, certain assumptions have to be made. First, we assume that the place of education of the production worker residing in state j is approximated by his state or region of birth. Secondly, we assume that the place of birth distribution for production workers is the same as that of the total native (i.e. U.S.) born population residing in the same jth state. The nature of the data require such assumptions, but they do not appear unrealistic.[1] As a result we have

[1] The data, which were available for this purpose, were for the place of birth of the total native born

$$u_{a_n} = P_{a_n} \Big/ \sum_{n=1}^{10} TN \tag{3}$$

where P denotes the population in a particular age class interval, a, and n denotes the region (one of the nine Census *regions*) or state of birth (education), where $n=1, \ldots, 10$.[2] Therefore, u is a percentage distribution of production workers according to their estimated place of education. We obtain the estimated number of production workers by industry and state in a particular age class interval who were educated in one of ten places by

$$\hat{L}_{a_n} = u_{a_n} \sum_{n=1}^{10} LCM \tag{4}$$

The result of procedures (1)–(4) is to generate a distribution of production workers by age, race, sex and place of education. All that remains is to determine each group's (cell's) median school years completed and the school expenditures that they received. *Population* median school years of the entire state population (by age, race and sex) were used to estimate the median school years of production workers.[3] For school expenditures per enrolled pupil I used data from *Statistical Abstracts*.[4] The human capital[5] manifest in the production worker is defined as the educational expenditures over the period which began t+7 (for simplicity it was assumed that education was uniformly begun at age 7) years after birth and terminated at the estimated median school years completed.[6] Thus,

$$\hat{H}C_{a_n} = \sum_{t+7}^{\hat{M}} E_{a_n} \cdot \hat{L}_{a_n} \tag{5}$$

where $\hat{H}C$ denotes the estimated human educational capital implicit in $\hat{L}$, E denotes the educational expenditures per pupil in the region or state, and $\sum_{t+7}^{\hat{M}}$ indicates the period over which the educational expenditures occurred.

Consequently, total human capital for white male production workers in industry i in state j may be written as

population for each state by age class intervals. More specifically, the data revealed the population by age class intervals residing in state j in 1960, who were born in state j, and the population residing in state j in 1960, who were born outside of state j. The latter category was subdivided into those who were born in one of the nine Census regions. Therefore, ten places of education were possible.

[2] Because data for median school years completed and school expenditures per enrolled pupil were not available for persons born outside the Continental United States, and would not be comparable even if the data were available, this group was eliminated from the computations. The result was to understate human capital by an amount equal to the human capital manifest in immigrant production workers.

[3] The result of this procedure was to assume that the median school years completed for production workers in each age class interval was constant in all of the industries in the state. This was obviously not the case. In reality, industries that employ production workers with low skills, e.g., lumber and wood products, were probably employing workers in the lower quartile of median school years completed for each age class interval, while industries that employed production workers with high skills, e.g., printing and publishing, were probably employing workers in the upper quartile. However, while this problem might be a serious one in explaining interindustry wage differentials, it should not be a problem in the explanation of interstate wage differentials. The feeling was that the distribution of workers of a certain educational level would be similar among the industries in each of the states; e.g., if printing and publishing generally employed workers with more median school years, this pattern would be similar among the states. The concern in this study was with variations in the variable from state to state within a two-digit industry, not with variations from industry to industry.

[4] Educational expenditure in historical dollars were converted to constant dollars using the *BLS* Consumer Price Index, which was available back to 1913. Prior to 1913, the Federal Reserve Consumer Price Index was used. There was no government expenditure price index that could be used.

[5] Some problems are implicit in the data. First, there are variations in the length of the school year among the states. This problem was particularly severe in the pre-World War II period and in the South. Since expenditures are related to the length of the school year, the problem was not particularly serious. Secondly, lower expenditures on nonwhite pupils is not taken into account because of data limitations. At worst, these two factors may slightly overstate southern school expenditures per pupil.

[6] Educational expenditures per pupil are assumed to incorporate a quality measure. The specification of a quality variable such as teacher's salaries, which in itself may not be a suitable quality variable to begin with, introduces severe collinearity problems. See Welch [18].

$$
(6) \qquad \hat{H}C_{ij} = \begin{bmatrix} \left(\sum_{i=7}^{\hat{M}} E_{11}\cdot\hat{L}_{11}\right) + \left(\sum_{i=7}^{\hat{M}} E_{12}\cdot\hat{L}_{12}\right) +, \cdots, + \left(\sum_{i=7}^{\hat{M}} E_{1.10}\cdot\hat{L}_{1.10}\right) \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ \left(\sum_{i=7}^{\hat{M}} E_{91}\cdot\hat{L}_{91}\right) + \left(\sum_{i=7}^{\hat{M}} E_{92}\cdot\hat{L}_{92}\right) +, \cdots, + \left(\sum_{i=7}^{\hat{M}} E_{9.10}\cdot\hat{L}_{9.10}\right) \end{bmatrix}
$$

Human educational capital for nonwhite males and for white and nonwhite females is calculated in an identical manner, and these are then summed and divided by total production worker manhours to produce an estimate of the human capital labor ratios.

REFERENCES

1. G. S. BECKER, *Human Capital*, New York 1964.
2. ———, *The Economics of Discrimination*, Chicago 1957.
3. J. W. BLOCK, "Regional Wage Differentials, 1907–1946," *Mon. Lab. Rev.*, Apr. 1948, *66*, 371–77.
4. G. H. BORTS, "The Equilization of Returns and Regional Economic Growth," *Amer. Econ. Rev.*, June 1960, *50*, 319–47.
5. H. M. DOUTY, "Wage Differentials: Forces and Counterforces," *Mon. Lab. Rev.*, Mar. 1968, *91*, 74–81.
6. D. E. FARRAR AND R. R. GLAUBER, "Multicollinearity in Regression Analysis: The Problem Revisited," *Rev. Econ. Statist.*, Feb. 1967, *49*, 92–103.
7. V. R. FUCHS AND R. PERLMAN, "Recent Trends in Southern Wage Differentials," *Rev. Econ. Statist.*, Aug. 1960, *42*, 292–300.
8. L. E. GALLAWAY, "The North-South Wage Differential," *Rev. Econ. Statist.*, Aug. 1963, *45*, 264–72.
9. ———, "The Aged and the Extent of Poverty in the United States," *Southern Econ. J.* Oct. 1966, *33*, 212–22.
10. ———, *Geographic Labor Mobility in the United States*, Social Security Administration, forthcoming, Washington.
11. ——— and G. W. SCULLY, "Minority Group Discrimination in the United States," Unpublished manuscript, Aug. 1968.
12. F. A. HANNA, "Contributions of Manufacturing Wages to Regional Differences in Per Capita Income," *Rev. Econ. Statist.*, Feb. 1951, *33*, 18–28.
13. J. W. KENDRICK, *Productivity Trends in United States*, Princeton 1961.
14. E. MUELLER, "Negro-White Differentials in Geographic Mobility," *Social Security Bulletin*, May 1967, Washington.
15. T. W. SCHULTZ, "Investment in Human Capital," *Amer. Econ. Rev.*, Mar. 1961, *51*, 1–17.
16. G. W. SCULLY, *Interstate Wage Differentials: A Cross Analysis*, unpublished doctoral thesis, Rutgers Univ., 1968.
17. M. SEGAL, "Regional Wage Differentials in Manufacturing in the Postwar Period," *Rev. Econ. Statist.*, May 1961, *43*, 248–55.
18. F. Welch, "Measurement of the Quality of Schooling," *Amer. Econ. Rev. Proc.*, May 1966, *56*, 379–92,
19. U.S. BUREAU of the CENSUS, *U.S. Census of Manufactures: 1958*, Washington 1961.
20. ———, *Census of the United States: 1960 Population*, Washington.
21. ———, *Statistical Abstract of the United States*, Washington 1903–1961.
22. U.S. DEPARTMENT of COMMERCE, *Historical Statistics of the United States: Colonial Times to 1957*, Washington 1961.
23. ———, *Historical Statistics of the United States: Continuation to 1962 and Revisions*, Washington 1965.
24. U. S. DEPARTMENT of LABOR, *Analysis of Work Stoppages*, Washington 1956–1960.

Long-Run Scale Adjustments of a Perfectly Competitive Firm and Industry

By C. E. FERGUSON AND THOMAS R. SAVING*

It has been well over forty-five years since, according to George Stigler, [8, p. 11]: "The concept of perfect competition received its complete formulation in Frank Knight's *Risk, Uncertainty and Profit* (1921)." Yet the age of a theory does not imply anything about its state of development. Most aspects of the theory of perfect competition are, of course, well known. But here again it is appropriate to quote Stigler: [8, p. 14]:

> The complete theory of competition cannot be known because it is an open-ended theory; it is always possible that a new range of problems will be posed in the framework, and then, no matter how well developed the theory was with respect to the earlier range of problems, it may require extensive elaboration in respects which previously it glossed over or ignored.

One such range of problems concerns the long-run scale adjustments of perfectly competitive firms and of the industry they comprise. More specifically, in the theory of perfect competition, the long-run responsiveness of firm and industry output to a change in factor price has been largely ignored. It has been shown that in the short run, the equilibrium output of a firm varies inversely (directly) with the price of a normal (inferior) factor [2], [4]. But this is all.

The chief object of this paper is to prove the following propositions for perfectly competitive firms and industries: (a) a firm's short- and long-run output adjustment to a change in factor price depends upon the expenditure elasticity of the factor, not upon its normality or inferiority; (b) regardless of the change in marginal cost, average cost must vary directly with factor price; (c) hence the long-run supply price of a perfectly competitive industry must vary directly, and its output inversely, with factor price.

As subsidiary results, we show that (a) from the production side, the relation between supply and expenditure elasticity is exactly analogous to the relation between demand and income elasticity in the theory of consumer behavior; (b) derived input demand functions have substitution and expenditure effects that are analogous to the substitution and income effects in the theory of demand;[1] and (c) long-run average cost is always inelastic with respect to factor price. The last two propositions are proved in the Appendix, to which no textual reference is made.

I

The basic model has been developed in a different context by Jacob Mosak [5] and

* The authors are professors of economics at Texas A&M University. The research embodied in this paper was partly supported by the TAMU Research Foundation and partly by the National Science Foundation, Grant GS-2430. After the manuscript was prepared, an article by Trout Rader [6] appeared containing some speculations about the likely effect of an increase in factor price upon equilibrium output and factor usage. In Section II, these results are made explicit.

[1] It should be noted, however, that derived input demand functions are always negatively sloped except for a curiosum shown in the Appendix. See [5], [7], and [3].

the present authors.[2] Consider a firm in a perfectly competitive industry. Assume that the industry is perfectly competitive in *all* factor markets and that factor supplies are perfectly elastic. The production function of the firm is

$$(1) \qquad q = f(x_1, x_2, \ldots, x_n),$$

where q is the rate of physical output and x_i is the rate of use of the i-th input. Over the relevant range all marginal products are positive and monotonically decreasing, as required by (4) and (6) below (i.e., $f_i > 0$ and $f_{ii} < 0$ for all i).

Denoting the price of the i-th input by p_i and product price by p, cost and profit are given, respectively, by

$$(2) \qquad c = \sum_{i=1}^{n} p_i x_i,$$

and

$$(3) \qquad \pi = pq - c.$$

Since our interest is focused on long-run adjustments, all factors are allowed to be variable. Thus we assume that the firm maximizes (3) subject to (1), (2), and to the condition for long-run competitive equilibrium (i.e., price equals marginal cost equals minimum long-run average cost). The well-known requirements for a unique maximum are

$$(4) \quad f_i - \lambda p_i = 0, \quad (i = 1, 2, \ldots, n)$$

$$(5) \qquad \frac{1}{\lambda} = p,$$

$$(6) \qquad \sum_i \sum_j f_{ij} dx_i dx_j < 0.$$

Inequality (6) must hold both by itself and subject to the linear constraint $\sum f_j dx_j = 0$. In the following, we denote the matrices associated with the unconstrained and constrained quadratics by [F*] and [F] respectively. Finally, to establish long-run industry equilibrium we require, in addition to (4)−(6), that marginal cost equal average cost, i.e.,

$$(7) \qquad \frac{1}{\lambda} = \frac{\sum_{j=1}^{n} p_j x_j}{f(x)}.$$

In what follows, we shall find it useful to relate the negative definite matrices in (6). To this end denote the determinants by F and F^* and the corresponding cofactors by subscripts, where the rows and columns of F are numbered $0, 1, \cdots, n$ and of F^*, $1, 2, \cdots, n$. It is then easily shown that

$$(8) \qquad F_{00} = F^*, \; F_{0i} = -\sum_{j=1}^{n} f_j F_{ij}^*. \qquad (i = 1, 2, \ldots, n)$$

Without loss of generality we may assume that there are an even number of inputs.[3] Thus, (6) and (8) imply that

$$(9) \qquad \frac{F_{ii}}{F} < 0, \quad (i = 1, 2, \ldots, n)$$

and

$$(10) \qquad \frac{F_{00}}{F} = \frac{F^*}{F} > 0.$$

The response of the long-run optimal scale of a firm to a change in factor prices can be derived by taking the total differential of (4) subject to condition (7). This is equivalent to forcing the response to changes in factor price to comply with cost minimization while maintaining long-run equilibrium. Note that price is still a

[2] See fn. 1.

[3] Some readers have questioned the validity of this assumption. Since [F] and [F*] are negative definite and their determinants and cofactors are always used in ratio form, the assumption cannot affect the sign of the ratio. This is all that matters. Next, F^*_{ij} is used in the expression $\Sigma f_j F^*_{ij}$ to define an inferior factor. Again the assumption is immaterial because by definition, a factor is inferior if the sum is positive and the number of factors is even, or if the sum is negative and the number of factors is odd. See [2] and [4].

parameter to the firm. In fact, in the model of long-run industry equilibrium, the entry or exit of firms forces price to equal the common minimum long-run average cost of all firms. Thus it is not necessary to take commodity price into account explicitly.

From (4) and (7) the relevant system of equations is

$$\sum_{j=1}^{n} f_{ij}dx_j - p_i d\lambda - \lambda dp_i = 0, \qquad (i = 1, 2, \ldots, n) \tag{11}$$
$$\sum_{j=1}^{n} f_j dx_j - \lambda\left[\sum_{j=1}^{n} p_j dx_j + \sum_{j=1}^{n} x_j dp_j\right] - c d\lambda = 0.$$

Without loss of generality, we may evaluate system (11) for the particular vector of factor prices given by column $[dp_1, 0, \cdots, 0]$. On this assumption, substituting (4) in (11) reduces the system to

$$-c d\lambda = \lambda x_1 dp_1, \tag{12}$$
$$\frac{-1}{\lambda} f_1 d\lambda + \sum_{j=1}^{n} f_{1j} dx_j = \lambda dp_1,$$
$$\frac{-1}{\lambda} f_i d\lambda + \sum_{j=1}^{n} f_{ij} dx_j = 0. \quad (i = 2, 3, \ldots, n)$$

The matrix of coefficients of (12) is

$$[H] = \begin{bmatrix} -c & 0 \\ \dfrac{-1}{\lambda} f_i & f_{ij} \end{bmatrix} \qquad (i, j = 1, 2, \ldots, n) \tag{13}$$

The determinants and cofactors of $[H]$ are related to F and F^* in the following way:

$$H = -cF_{00} = -cF^*, \tag{14}$$
$$H_{0i} = \frac{-1}{\lambda} F_{0i} = \frac{1}{\lambda} \sum_{j=1}^{n} f_j F^*_{ij}, \qquad (i = 1, 2, \ldots, n)$$
$$H_{i0} = 0, \qquad (i = 1, 2, \ldots, n)$$
$$H_{ij} = -cF^*_{ij}. \quad (i, j = 1, 2, \ldots, n)$$

Solving (12) by Cramer's Rule and using (14) we obtain

$$dx_j = \left(x_1 \frac{F_{0j}}{cF^*} + \lambda \frac{F^*_{1j}}{F^*}\right) dp_1, \qquad (j = 1, 2, \ldots, n) \tag{15}$$
$$d\lambda = -\frac{\lambda x_1}{c} dp_1.$$

From (1) the change in total output is

$$dq = \sum_{j=1}^{n} f_j dx_j. \tag{16}$$

Substituting for the dx_j from (15) and noting that $F = \sum_j f_j F_{0j}$, we obtain the expression for the change in long-run optimal scale in response to a change in factor price:

$$\frac{dq}{dp_1} = \frac{x_1 F}{cF^*} + \frac{\lambda}{F^*} \sum_{j=1}^{n} f_j F^*_{1j}. \tag{17}$$

The first term on the right-hand side of (17) is always positive inasmuch as each element in the expression is positive. Next, since λ and F^* are positive, the second term is positive or negative according as $\sum_j f_j F^*_{1j} \gtrless 0$. Now differentiate (4) and (2) and solve the resulting system of $n+1$ equations for dx_1/dc to obtain

$$\frac{dx_1}{dc} = -\frac{\lambda}{F} \sum_{j=1}^{n} f_j F^*_{1j}. \tag{18}$$

From (18) we see that the crucial part of the second term on the right-hand side of (17) will be positive if, and only if, the factor under consideration is inferior.[4] Thus for an inferior factor, the long-run optimal scale of the firm must increase when factor price increases. But we can say considerably more than this.

Define the expenditure elasticity of a

[4] An inferior factor is one whose use declines when output (and thus cost) is expanded at constant factor prices. Alternatively, an inferior factor is one whose expenditure elasticity is negative.

factor as the relative responsiveness of factor usage to a change in total expenditure on all resources. That is, the expenditure elasticity is the proportional change in the use of a given factor divided by the proportional change in expenditure on all factors. Denoting the expenditure elasticity of factor 1 by n_{1c}, we have, from the definition and (18),

$$\eta_{1c} = \left(\frac{dx_1}{dc}\,\frac{c}{x_1}\right) = -\frac{\lambda c}{x_1 F}\sum_{j=1}^{n} f_j F_{1j}^{*}. \tag{19}$$

Solving (19) we have

$$\sum_{j=1}^{n} f_j F_{1j}^{*} = -\frac{x_1 F}{\lambda c}\,\eta_{1c}. \tag{20}$$

Substituting (20) into (17) yields the crucial relation:

$$\frac{dq}{dp_1} = \frac{x_1 F}{cF^{*}}(1 - \eta_{1c}). \tag{21}$$

It was noted above that x_1, c, F, and F^{*} are all positive. Hence the effect of a change in factor price on the long-run optimal scale of the firm depends exclusively upon the expenditure elasticity of the input whose price has changed. In particular, if a factor has a unitary expenditure elasticity, a change in its price will not affect the long-run optimal scale of firms. Otherwise, the long-run optimal scale will increase or decrease according as the expenditure elasticity is less or greater than unity.

Let us now introduce the following classification: factor 1 is inferior if $\eta_{1c}<0$, normal if $0<\eta_{1c}<1$, and superior if $\eta_{1c}>1$. Thus for normal and inferior factors, an increase in factor price will result in an increase in the long-run optimal scale of the firm. On the other hand, an increase in the price of a superior factor will result in a decrease in the long-run optimal scale of the firm. These relations are illustrated in Figure 2; but a discussion of the figure is deferred until we have proved that long-run average cost must vary directly with factor price, irrespective of factor classification.

The discussion in the paragraph above is in terms of an increase in factor price. Generalizing, the long-run equilibrium output of a perfectly competitive firm must always vary directly with the price of an inferior or normal input and must always vary inversely with the price of a superior input. This is a result of some surprise inasmuch as it has customarily been thought that equilibrium output varies inversely with factor price (except for inferior factors). Nonetheless, the explanation of this phenomenon is relatively straightforward once the relation between the expenditure elasticity of a factor and the change in its relative share is taken into account.

Let κ_1 denote the proportion of total cost accounted for by factor 1. In the equilibria with which we deal this is also, given competitive imputations, the relative distributional share of factor 1. The change in the factor share as expenditure and output increase is

$$\frac{d\kappa_1}{dc} = \frac{\kappa_1}{c}(\eta_{1c} - 1). \tag{22}$$

As (22) clearly shows, the share of a superior factor ($\eta_{1c}>1$) increases as expenditure and output increase, whereas the share of an inferior or normal factor ($\eta_{1c}<1$) decreases. Thus when a factor price increases, there will be an increase in output and a decrease in the relative share of the now more expensive factor if that factor is inferior or normal; the opposite relation holds for a superior factor.

II

It has been shown that an increase in the price of an inferior or normal factor will lead to an increase in the long-run optimal scale of an individual firm in a competitive industry. It does not follow, however, that

long-run equilibrium industry output increases as well. In fact, our intuition and simple logic assure us that this is not the case. An increase in factor price must surely increase long-run average cost for every level of output; therefore long-run supply price for the industry must increase and equilibrium industry output must decline. Since the optimal scale of the firm must increase if the factor whose price has increased has an expenditure elasticity less than one, i.e., is inferior or normal, the reduction in industry output must be accomplished by means of an exodus of firms from the industry.

To establish this proposition in a simple way,[5] consider any point (AC°, q°) on the long-run average cost curve. Let one factor price change (all others remaining constant) and consider the corresponding point (AC^{1}, q°) on the new long-run average cost curve. Clearly the cost of the new input set at the new prices is higher than the same set purchased at the old factor prices. But at the old factor prices, the firm chose a different set of inputs, indicating that the cost of purchasing the old set was less than the cost of the new set at the old prices. Thus the cost of the new set of inputs at the new prices exceeds the cost of the new set at the old prices, which in turn exceeds the cost of the old set at the old prices.

Therefore, average cost must have increased, i.e., $AC^{1} > AC^{\circ}$ at output q°. But since the choice of q° was arbitrary, long-run average cost must be higher at every output. Now if output q° corresponded to the minimum point on the original long-run average cost curve, then every point on the long-run average cost curve associated with the new set of factor prices must be greater than AC°. Hence the new level of minimum average cost must be higher so that the level of minimum average cost varies directly with factor price irrespective of factor classification.

To summarize, whether factor 1 is inferior, normal, or superior, an increase in its price always shifts long-run total cost upward for any level of output. Consequently, whenever a factor price increases, the long-run average cost curve is displaced upward, with its new minimum point corresponding to a larger or a smaller output according as the expenditure elasticity of the factor whose price changes is less than or greater than unity. Therefore, long-run industry price must rise and, given the demand function, long-run equilibrium output must decline.

Let us emphasize again that this result is independent of the classification of the factor whose price changes. But note that the short-run industry response is not independent of factor inferiority. As shown by Bear [2] (and our equation (1.5) in the Appendix), both long-run and short-run marginal cost shift downward in response to an increase in the price of an inferior factor. Thus short-run industry output increases when the price of an inferior factor increases. But as shown above, long-run industry output must decline. This occurs, of course, because in the short-run firms in this industry earn losses; there is accordingly an exodus of firms from the industry, and this continues until price is equal to the now higher minimum average cost.

III

The chief results of previous sections are summarized graphically in this section. The three panels of Figure 1 each show initial long-run average and marginal cost curves (AC_0, MC_0) and the same curves after they have been displaced by an increase in the price of some factor (AC_1, MC_1). If the factor in question is inferior, Panel a, marginal cost shifts downward to

[5] Mathematical derivation is given in Note 1 in the Appendix.

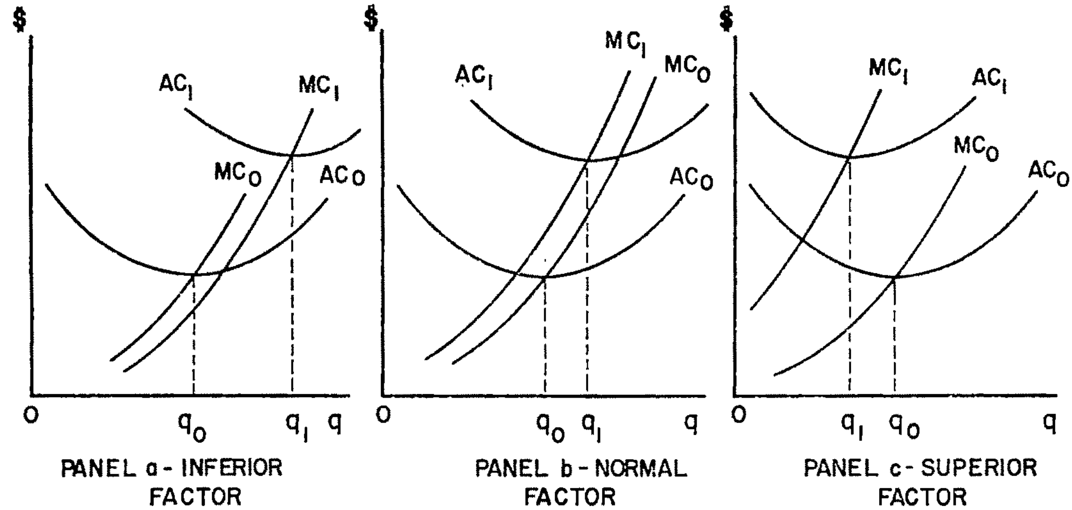


FIGURE 1 - SCHEMATIC ILLUSTRATION OF SHIFTS IN AC AND MC WHEN FACTOR PRICE INCREASES

the right, average cost shifts upward to the right, and minimum average cost occurs at a greater rate of output.

If the factor is normal, marginal cost shifts upward to the left, average cost shifts upward to the right, and minimum average cost corresponds to a larger volume of output. Finally, if the input is superior, both marginal and average cost shift upward to the left, and minimum average cost occurs at a smaller rate of output. Since long-run supply price must increase in all cases, the reduction in industry output must be accomplished by an exodus of firms if the factor whose price rises is normal or inferior.

The displacement of minimum long-run average cost is illustrated more clearly in Figure 2. AC_0 and MC_0 are the initial long-run average and marginal cost curves. We again assume that the price of a factor increases. The areas and lines with the η_{1c} labels indicate the direction in which minimum average cost is displaced. The three cases discussed above are quite clear.[6] Two interesting additional cases emerge in which the expenditure elasticity of the factor whose price increases is zero or one. The following propositions are proved in Note 1 of the Appendix.

[6] That is, the cases of factor superiority ($\eta_{1c}>1$), factor normality ($0<\eta_{1c}<1$), and factor inferiority ($\eta_{1c}<0$).

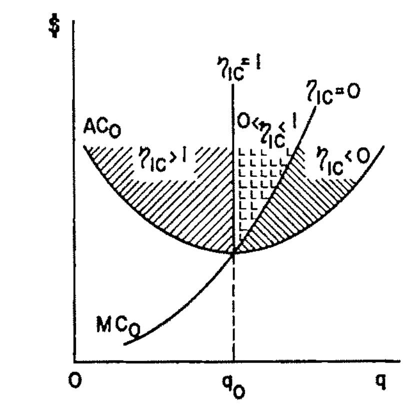


FIGURE 2 - SHIFTS IN MINIMUM AC AND THE EXPENDITURE ELASTICITY

In the first situation (expenditure elasticity of zero), the long-run average cost curve must move upward along the original marginal cost curve. Since $\eta_{1e}=0$ by hypothesis, a change in total expenditure does not affect the usage of the factor whose price increased. Hence marginal cost does not change; and the new average cost curve must intersect the unchanged marginal cost curve at the point where the former attains its minimum value.

This relation has a straightforward economic interpretation. A factor whose expenditure elasticity is zero is economically fixed even though it is technologically variable. That is, profit-maximizing entrepreneurs, who can vary the factor in question, find that their long-run optimal adjustment dictates using the same amount of the factor at every level of output. Therefore the factor is economically fixed; and changes in its price are tantamount to changes in fixed cost, which do not affect marginal cost. In the second situation (expenditure elasticity of one), the long-run average cost curve simply shifts vertically and the minimum average cost output remains unchanged.

IV

In competitive industry, long-run equilibrium price will always vary directly with factor price, while long-run equilibrium output will vary inversely. It is only within the individual firms that the long-run adjustment may be different; and we have seen that the nature of these adjustments depends critically upon the expenditure elasticity of the input whose price changes. Using the classification scheme introduced above, we may summarize the results as follows: given a change in input price, (a) if the factor is inferior, long-run marginal cost varies inversely and long-run equilibrium output varies directly with factor price; (b) if the factor is normal, both long-run marginal cost and equilibrium output vary directly with factor price; and (c) if the factor is superior, long-run marginal cost varies directly and equilibrium output inversely with factor price.

Appendix

Note 1. *Displacement of Long Run Average Cost.* In the text we argue that long-run average cost must vary directly with factor price. This result can be easily confirmed by differentiating the first-order conditions (4) subject to the constraint that output remain constant, i.e., the differential of (1) be zero. Output is held constant so as to enable us to infer the behavior of average cost from the behavior of total cost. Performing the indicated operations yields the following system of equations:

$$\sum_{j=1}^{n} f_j dx_j = 0,$$

$$(1.1)\qquad -p_i d\lambda + \sum_{j=1}^{n} f_{ij} dx_j = \lambda dp_1.$$

$$(i = 1, 2, \ldots, n)$$

The matrix of coefficients of (1.1) is

$$(1.2)\qquad [G] = \begin{bmatrix} 0 & f_i \\ -p_j & f_{ij} \end{bmatrix}.$$

$$(i, j = 1, 2, \ldots, n)$$

Using (4) and (8) establishes the following relations among the determinants and cofactors of $[G]$, $[F]$, and $[F^*]$:

$$G = -\frac{1}{\lambda} F, \qquad G_{ij} = -\frac{1}{\lambda} F_{ij},$$

$$(1.3)\qquad (i = 0, 1, \cdots, n;\ j = 1, 2, \ldots, n)$$

$$G_{00} = F_{00} = F^*, \qquad G_{i0} = F_{i0}.$$

$$(i = 1, 2, \ldots, n)$$

Solving (1.1) by Cramer's Rule and using (1.3), one obtains

$$dx_j = \frac{\lambda}{F} \sum_{i=1}^{n} F_{ij} dp_i, \quad (j = 1, 2, \ldots, n)$$

$$(1.4)$$

$$d\lambda = -\frac{\lambda^2}{F} \sum_{i=1}^{n} F_{i0} dp_i.$$

Again without loss of generality we may assume that the price of input 1 is the only one that changes. Upon substitution of the factor-price vector column $(dp_1, 0, \cdots, 0)$, (1.4) becomes[7]

$$dx_j = \frac{\lambda}{F} F_{1j} dp_1, \quad (j = 1, 2, \ldots, n)$$

$$(1.5) \quad d\lambda = -\frac{\lambda^2}{F} F_{10} dp_1 = \frac{\lambda^2}{F} \sum_{j=1}^{n} f_j F_{1j}^* dp_1 = -\frac{x_1 \lambda}{c} \eta_{1c} dp_1.$$

First note that $(d\lambda/dp_1) \gtrless 0$ according as

$$\sum_{j=1}^{n} f_j F_{1j}^* \gtrless 0$$

since λ and F are positive. Thus λ varies directly or inversely with factor price according as the factor whose price changes is inferior or noninferior. But since λ is the reciprocal of long-run marginal cost, the latter shifts downward when the price of an inferior factor rises and upward when the price of a noninferior factor rises.[8]

The results obtained in (1.5) can now be used to show that long-run average cost must vary directly with factor price irrespective of factor classification. Differentiating the cost function (2) and using the factor-price vector column $(dp_1, 0, \ldots, 0)$, one obtains

$$(1.6) \quad dc = \sum_{j=1}^{n} p_j dx_j + x_1 dp_1. \qquad (dq = 0)$$

Substituting (1.5) in (1.6) yields

$$(1.7) \quad \frac{dc}{dp_1} = \frac{1}{F} \sum_{j=1}^{n} f_j F_{1j} + x_1. \quad (dq = 0)$$

But

$$\sum_{j=1}^{n} f_j F_{1j}$$

is an expansion of F by alien cofactors and is accordingly zero. Hence the change in long-run total cost resulting from a change in factor price, output constant, is

$$(1.8) \quad \frac{dc}{dp_1} = x_1 > 0. \qquad (dq = 0)$$

Since total cost increases for each quantity produced, it follows that long-run average cost increases as well.

Note 2. *Derived Input Demand Functions.* From (1.5) we immediately obtain the slope properties of the constant-output demand function for an input:

$$(2.1) \quad \frac{dx_j}{dp_j} = \frac{\lambda F_{1j}}{F} = \frac{1}{p} \frac{F_{1j}}{F}. \qquad (j = 1, 2, \ldots, n)$$

Now introduce the customary definition of the partial elasticity of substitution [1, p. 504]:

$$(2.2) \quad \sigma_{1j} = \frac{c}{x_1 x_j} \frac{F_{1j}}{F}.$$

Using (2.2) in (2.1), the constant-output demand curves may be given the following representation:

$$(2.3) \quad \left(\frac{dx_j}{dp_1}\right)_{dq=0} = \frac{-x_1 x_j}{pc} \sigma_{1j}. \qquad (j = 1, 2, \ldots, n)$$

Since $(F_{ii}/F) < 0$, $\sigma_{ii} < 0$. Therefore, all constant-output factor demand functions are negatively sloped with respect to their own price. By customary definition [1, p. 505], inputs 1 and j are said to be competitive or complementary according as $\sigma_{1j} \gtrless 0$. Thus output held constant, a rise in the price of input 1 will augment or diminish the use of input j according as the inputs are competitive or complementary. It should further be noted that (a) this result is independent of factor classification and (b) the change represents a pure substitution effect inasmuch as the movements are restricted to an original isoquant.

The slope characteristics of unrestricted input demand functions may be derived by

[7] The proof that marginal cost is unchanged when the expenditure elasticity of the factor is zero follows from the second equation in (1.5).

[8] This result was shown by Bear [2].

displacing (4) subject to the constraint imposed by the production function:

$$
\begin{aligned}
&\sum_{j=1}^{n} f_j dx_j = dq,\\
(2.4)\qquad &-p_1 d\lambda + \sum_{j=1}^{n} f_{1j} dx_j = \lambda dp_1,\\
&-p_i d\lambda + \sum_{j=1}^{n} f_{ij} dx_j = 0.\\
&\qquad\qquad (i = 2, 3, \ldots, n)
\end{aligned}
$$

The matrix of coefficients is $[G]$ and has the properties shown in (1.3) above. Solving (2.4) by Cramer's Rule yields

$$(2.5)\qquad d\lambda = -\frac{\lambda F^*}{F} dq + \lambda^2 \frac{\sum_{j=1}^{n} f_j F_{1j}^* dp_1}{F},$$

$$(2.6)\qquad dx_j = \frac{\sum_{k=1}^{n} f_k F_{jk}^*}{F} dq + \frac{\lambda F_{1j}}{F} dp_1. \qquad (j = 1, 2, \ldots, n)$$

These equations hold for a single profit-maximizing firm in a perfectly competitive industry. Two distinct cases may be discussed. First, in the short run, price is a parametrically given constant equal to marginal cost. Since λ is the reciprocal of marginal cost, $dp=0$ implies $d\lambda=0$. Using this information in (2.5) and substituting (19), (2.2), and (2.5) in (2.6) yields the derived short-run input demand functions whose characteristics we wish to determine:

$$(2.7)\qquad \frac{dx_j}{dp_1} = -\frac{x_1 x_j F}{\lambda c^2 F^*} \eta_{1c}\eta_{jc} + \frac{\lambda x_1 x_j}{c} \sigma_{1j}. \qquad (j = 1, 2, \ldots, n)$$

The second term on the right-hand side of (2.7) is precisely the substitution effect we isolated when analyzing the constant-output demand functions. All statements made in that connection apply here as well. The first term on the right-hand side of (2.7) is the expenditure effect that reflects the change in output and expenditure incident to profit-maximizing adjustments to a change in factor price.

Let us examine the change for factor 1:

$$(2.8)\qquad \frac{dx_1}{dp_1} = -\frac{x_1^2 F}{\lambda c^2 F^*} (\eta_{1c})^2 + \frac{\lambda x_1^2}{c} \sigma_{11}.$$

Both terms on the right-hand side of (2.8) are necessarily negative, so quantity demanded varies inversely with price irrespective of the factor classification. The expenditure elasticity enters the expenditure effect as a squared term. Thus the expenditure effect in the theory of the firm can never have perverse effects as can the corresponding income elasticity in the theory of consumer behavior.

Second, in the long run, price must adjust so as to equal the changing minimum long-run average cost. Thus we can define a long-run equilibrium derived input demand function for each firm by requiring that long-run marginal cost equal long run, average cost. Consequently,

$$(2.9)\qquad d\lambda = -\frac{\lambda x_1}{c} dp_1.$$

Substituting (19), (2.2), (2.5) and (2.9) in (2.6) yields the result we seek:

$$(2.10)\qquad \frac{dx_1}{dp_1} = \frac{x_1^2 F}{\lambda c^2 F^*} (\eta_{1c} - \eta_{1c}^2) + \frac{\lambda x_1^2}{c} \sigma_{11}.$$

We can again separate the total effect of a change in factor price into a substitution effect and an expenditure effect. The substitution effect is, of course, necessarily negative; but in this case the expenditure effect may be positive. Thus in the long run an increase in factor price may result in an increase in the usage of the factor in question. This curiosum can only occur if the expenditure elasticity is between 0 and 1 since it is only in this case that the combination of a positive relation between output and both factor usage and factor price occurs (see Figure 1). In all other cases, the first term on the right-hand side of (2.10) is necessarily nonpositive (it is 0 when $\eta_{1c}=0$ or when $\eta_{1c}=1$). Note, however, that this result does

not apply to the industry demand function since industry output necessarily declines when a factor price rises. It should further be noted that this is not a derived demand function of the conventional form. But it is interesting to observe that in shifting from one long-run equilibrium to another, a firm's usage of a factor may vary directly with its price.

Note 3. *Responsiveness of Minimum Long-Run Average Cost to Factor Price Changes.* The results shown in (15) are derived assuming that the firm is operating at the minimum point on its average cost curve. Hence

$$AC = \frac{1}{\lambda}, \tag{3.1}$$

$$d(AC) = -\frac{d\lambda}{\lambda^2}. \tag{3.2}$$

Now substituting (15) into (3.2) yields

$$d(AC) = \frac{x_1}{\lambda c}\, dp_1, \tag{3.3}$$

so that the elasticity of the minimum long-run average cost with respect to factor price is

$$\frac{d(AC)}{dp_1}\,\frac{p_1}{AC} = \kappa_1, \tag{3.4}$$

where κ_1 is the relative factor share of factor 1. Since in a multifactor model $\kappa_1 < 1$, it follows that minimum long-run average cost must be inelastic with respect to factor price. This result is to be expected since even at constant output, an X percent increase in factor price must increase average cost by less than X percent. Since the new minimum average cost must be at least as low as the constant output average cost, it too must rise by less than X percent.

REFERENCES

1. R. G. D. ALLEN, *Mathematical Analysis for Economists*, London, 1938.
2. D. V. T. BEAR, "Inferior Inputs and the Firm," *J. Polit. Econ.*, June 1965, *73*, 287–89.
3. C. E. FERGUSON, "Production, Prices, and the Theory of Jointly-Derived Input Demand Functions," *Economica*, Nov. 1966, *33*, 454–61.
4. ———, "'Inferior Factors' and the Theories of Production and Input Demand," *Economica*, May 1968, *35*, 140–50.
5. J. L. MOSAK, "Interrelations of Production, Price, and Derived Demand," *J. Polit. Econ.*, Dec. 1938, *46*, 761–87.
6. T. RADER, "Normally, Factor Inputs are Never Gross Substitutes," *J. Polit. Econ.*, Jan. 1968, *76*, 38–43.
7. T. R. SAVING, "Note on Factor Demand Elasticity: The Competitive Case," *Econometrica*, July 1963, *31*, 555–57.
8. G. J. STIGLER, "Perfect Competition Historically Contemplated," *J. Polit. Econ.*, Feb. 1957, *65*, 1–17.

A Re-Examination of the Modigliani-Miller Theorem

By JOSEPH E. STIGLITZ*

In their classic paper of 1958, Franco Modigliani and Merton H. Miller demonstrated that the cost of capital for a firm was independent of the debt-equity ratio [13]. Although much of the subsequent discussion has focused on the realism of particular assumptions [3], [7], there have been few attempts to delineate exactly the class of assumptions under which the M-M theorem obtains.[1] In particular, five limitations of the M-M proof may be noted:

1. It depended on the existence of risk classes.
2. The use of risk classes seemed to imply objective rather than subjective probability distributions over the possible outcomes.
3. It was based on partial equilibrium rather than general equilibrium analysis.
4. It was not clear whether the theorem held only for competitive markets.
5. Except under special circumstances, it was not clear how the possibility of firm bankruptcy affected the validity of the theorem.

In Section I, we show in the context of a general equilibrium state preference model that the M-M theorem holds under much more general conditions than those assumed in their original study. The validity of the theorem does not depend on the existence of risk classes, on the competitiveness of the capital market, or on the agreement of individuals about the probability distribution of outcomes.[2]

The two assumptions which do appear to be important for our proof are (a) individuals can borrow at the same market rate of interest as firms and (b) there is no bankruptcy.[3] But it is these assumptions which appear to be the center of much of the criticism of the M-M analysis. In Section II, we show that the M-M results may still be valid even if there are limitations on individual borrowing, and in Section III, we show that the possibility of bankruptcy raises more serious problems, although the M-M theorem can still be shown to hold under somewhat more stringent conditions.

I. *The Basic Theorem*

Consider a firm whose gross returns, X (before paying bondholders but after paying all non-capital factors of production) are uncertain. We can consider X as a function of the state of the world θ. One dollar invested in a perfectly safe bond yields a gross return of r^*, so that r^*-1 is

* Associate professor, Cowles Foundation, Yale University. This is a revised version of Cowles Foundation Discussion Paper No. 242, presented at the December 1967 meetings of the Econometric Society. The research described in this paper was carried out under grants from the National Science Foundation and from the Ford Foundation. I am deeply indebted to D. Cass, A. Klevorick, M. Miller, and W. Nordhaus for extensive discussions on these problems and detailed comments on a previous draft.

[1] Exceptions are the work of Hirschleifer [9], [10] and Robichek and Meyers [19], who used the Arrow-Debreu model (which assumes at least as many securities as states of the world) and the doctoral dissertation of G. Pye [18]. More recently Sher [21] has concerned himself with some of the difficulties raised by bankruptcy. For other general equilibrium portfolio (stock-market) models, see Sharpe [20], Lintner [11], Mossin [17], and Diamond [6].

[2] Except that they must agree that there is zero probability of bankruptcy. See discussion in text.

[3] It should be clear that these assumptions are not completely independent. Presumably, one of the most important reasons individuals cannot borrow at the same rate as firms is that there is a higher probability of default.

the market rate of interest. If there is any chance of bankruptcy, the nominal rate $\hat{r}$ which the firm must pay on its bonds will depend on the number issued. If principal payments plus interest exceed gross profits, X, the firm goes bankrupt, and the gross profits are divided among the bondholders.[4] Thus the gross return on a dollar invested in the bonds of the firm depends on state θ

$$(1)\qquad \hat{r}(\theta) = \begin{cases} \hat{r} & \text{if } \hat{r}B \leq X(\theta) \\ \dfrac{X(\theta)}{B} & \text{if } \hat{r}B \geq X(\theta). \end{cases}$$

Earnings per dollar invested in equity in state θ are given by

$$(2)\quad e(\theta) = \begin{cases} [X(\theta) - \hat{r}B]/E & \text{if } \hat{r}B \leq X(\theta) \\ 0 & \text{if } \hat{r}B \geq X(\theta) \end{cases}$$

where E is the value of the firm's equity. The value of the firm is

$$(3)\qquad V = E + B.$$

Individuals will be assumed to evaluate alternative portfolios in terms of their income patterns across the states of nature.

We now prove the following proposition.

> Assume there is no bankruptcy and individuals can borrow and lend at the market rate of interest. If there exists a general equilibrium with each firm having a particular debt-equity ratio and a particular value, then there exists another general equilibrium solution for the economy with any firm having any other debt-equity ratio but with the value of all firms and the market rate of interest unchanged.

[4] Throughout the discussion, we limit ourselves to a two-period model. In a two-period model, a firm either makes its interest payments or goes bankrupt. In a multiperiod model, the firm can, in addition, defer the interest or principal payments. If there is a positive probability of such deferral, the market will force the firm to pay a higher nominal rate of interest. If there are large transaction costs involved in bankruptcy or deferral, the M-M theorem would not hold. Throughout the discussion we shall assume that there are no flotation costs and no taxes.

Proof: Let w^j be the j^{th} individual's wealth, $E^j{}_i$, the value of his shares of the i^{th} firm, B^j the number of bonds he owns.[5] Assume the i^{th} firm, whose value is V_i, issues B_i bonds. The j^{th} individual's budget constraint may be written

$$(4)\qquad w^j = \sum_i E_i^j + B^j.$$

If we let $\alpha^j{}_i = E^j{}_i/E_i$, the share of the i^{th} firm's equity owned by the j^{th} individual, (4) becomes

$$(5)\qquad w^j = \sum_i \alpha_i^j E_i + B^j.$$

Then his income in state θ may be written

$$(6)\qquad \begin{aligned} Y^j(\theta) &= \sum_{i=1}^{n} (X_i - r^* B_i)\alpha_i^j \\ &\quad + r^*\left(w^j - \sum_{i=1}^{n} \alpha_i^j (V_i - B_i)\right) \\ &= \sum_{i=1}^{n} X_i \alpha_i^j + r^*\left(w^j - \sum_{i=1}^{n} \alpha_i^j V_i\right). \end{aligned}$$

If, as B_i changes, V_i remains unchanged, the individual's opportunity set does not change, and the set of $\alpha^j{}_i$ which maximizes the individual's utility is unchanged. If

$$\sum_j \alpha^j{}_i = 1$$

before, i.e., demand for shares equalled supply of shares, it still does. The total net demand for bonds is

$$\sum_j \left(w^j - \sum_i \alpha_i^j (V_i - B_i)\right) + \sum_i B_i = \sum_j w^j - \sum_i V_i.$$

If the market was in equilibrium initially,

$$\sum_j w^j - \sum_i V_i = 0,$$

i.e., excess demand equalled zero. If as the debt equity ratio changes, all V_i remain unchanged, excess demand remains at zero.

[5] By convention, one bond costs one dollar.

An alternative way of seeing this is the following. We may rewrite (6) as

$$(6') \quad Y^j(\theta) = \sum_i e_i(\theta)E_i^j + r^*\left(w^j - \sum_{i=1}^n E_i^j\right).$$

Assume now that the first firm, say, issues no bonds. If we let carets denote the values of the various variables in this situation, the opportunity set is given by

$$(6'') \quad \hat{Y}^j(\theta) = \sum_i \hat{e}_i(\theta)\hat{E}_i^j + \hat{r}^*\left(w^j - \sum_i \hat{E}_i^j\right).$$

Assume $r^*=\hat{r}^*$, $E_i=\hat{E}_i$, $i\geq 2$. Then from (2), $e_i(\theta)=\hat{e}_i(\theta)$, $i\geq 2$. If $\hat{E}_1=E_1+B_1$, then the opportunity sets described by (6′) and (6″) are identical. To see this, assume that for each dollar of equity he owned in the first firm in the initial situation, the individual borrows B_1/E_1 in addition to B^j

$$\text{so } \hat{B}^j = B^j + E_1^j\frac{B_1}{E_1}.$$

With the proceeds of the loan he increases his holdings of equities in the first firm, so

$$(7) \qquad \hat{E}_1^j = E_1^j + E_1^j\frac{B_1}{E_1} = E_1^j\left(\frac{V_1}{E_1}\right).$$

His income in state θ is then given by

$$(8) \qquad \begin{aligned} \hat{Y}^j(\theta) &= \frac{X_1E_1^j}{E_1} + \sum_{i=2}^n e_i(\theta)E_i^j \\ &\quad + r^*\left(w^j - \sum_{i=2}^n E_i^j - \frac{E_1^jV_1}{E_1}\right) \\ &= \left(\frac{X_1 - r^*B_1}{E_1}\right)E_1^j + \sum_{i=2}^n e_i(\theta)E_i^j \\ &\quad + r^*\left(w^j - \sum_{i=1}^n E_i^j\right) \end{aligned}$$

which is identical to (6′).

Since his opportunity set has not been changed as a result of the change in the debt-equity ratio of the firm, if he was maximizing his utility in the initial situation, the optimal allocation in the new situation is identical to that in the initial situation with the one modification given above.

We now need to show that the markets for the firm's equities and the market for bonds will clear. Summing (7) over all individuals, we obtain

$$\sum_j \hat{E}_1^j = \frac{V_1}{E_1}\sum_j E_1^j.$$

Thus the demand for equities has increased by a factor V_1/E_1. But since $\hat{E}_1/E_1 = V_1/E_1$, the supply has increased by exactly the same proportion, so if demand equalled supply before it also does now. Similarly, the increase in the demand for bonds by individuals equals $(B_1/E_1)\sum E_1^j = B_1$. But this exactly equals the decrease in the demand for bonds by the first firm.

It should be emphasized that in this proof, $X(\theta)$ is subjectively determined; moreover no assumptions about the size of firms, the source of the uncertainty, and the existence of risk classes have been made. The only restriction on the individual's behavior is that he evaluates alternative portfolios in terms of the income stream they generate. The two crucial assumptions were (a) all individuals agree that for all firms $X_i(\theta) > r^*B$ for all θ (see Section III); and (b) individuals can borrow and lend at the market rate of interest. This assumption is considerably weaker than the assumption of a competitive capital market, since no assumption about the number of firms has been made: the market rate of interest need not be invariant to the supply of bonds by any single firm.

II. *Limitations on Individual Borrowing*

One of the main objections raised to the M-M analysis is that individuals cannot borrow at the same rate of interest as firms. First, it should be noted (see [13])

that the analysis does not require that individuals actually borrow from the market, but only that they change their holdings of bonds. A problem can arise then only if an individual has *no* bonds in his portfolio.

Although the requirement that all individuals hold bonds does place restrictions on the possible debt-equity ratios of different firms, there still need not be an optimal debt-equity ratio for any single firm. Assume we have some general equilibrium situation where $B^j \geq 0$ for all j. Then so long as B_i satisfy the inequalities

$$(9) \qquad \sum_i \alpha_i^j B_i \geq w^j - \sum_i \alpha_i^j V_i \qquad \text{for all } j$$

all individuals will be lenders. If there were two firms, the constraints (9) would imply that (B_1, B_2) lie in the shaded area shown in Figure 1. For any pair of (B_1, B_2) in the region, there will exist a general equilibrium in which the values of both firms are identical to that in the original situation.

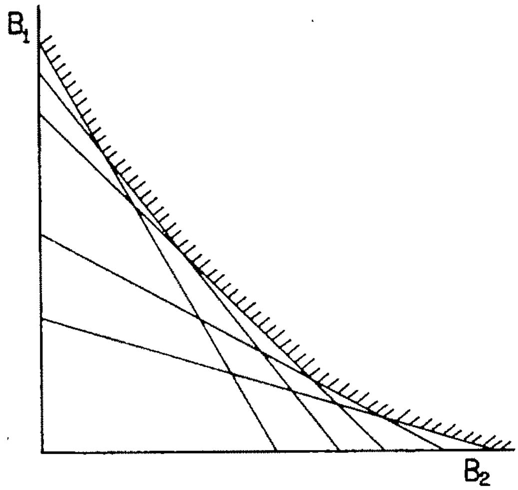


FIGURE 1

So far, none of our results have depended on the existence of risk classes.[6] The following two results depend on more than one firm having the same pattern of returns across the states of nature.

We shall first show that if there are two (or more) firms with the same pattern of returns and individuals can sell short, then the two firms must have the same value, independent of the debt-equity ratio.

We follow M-M in assuming for simplicity that one of the two firms has no outstanding debt, so $V_1 = E_1$. The second firm issues B_2 bonds, so $V_2 = B_2 + E_2$.

Consider first an individual who owns α_1 of the shares of the first firm, yielding an income pattern $\alpha_1 X_1(\theta)$. If instead he purchases α_1 of the shares of the second firm, at a cost of $\alpha_1 E_2$, and buys $\alpha_1 B_2$ bonds, his income in state θ is $\alpha_1(X_2(\theta) - r^* B_2) + \alpha_1 r^* B_2 = \alpha_1 X_2(\theta)$ which is identical to his income in state θ in the previous situation. But the cost of purchasing α_1 of the shares of the first company is $\alpha_1 V_1$ which is *greater* than $\alpha_1(E_2 + B_2) = \alpha_1 V_2$ if $V_1 > V_2$. Accordingly, if V_1 were greater than V_2, all holders of shares in the first company would sell their shares and purchase shares in the second firm, driving the value of the second firm up and that of the first down. Now consider an individual who wishes to lend money. If he sells short α_2 of the shares of the second firm and buys α_2 of the shares of the first firm, he receives a perfectly safe return of[7] $-\alpha_2(X_2 - r^* B_2) + \alpha_2 X_1 = \alpha_2 r^* B_2$ at a net cost of $-\alpha_2(V_2 - B_2) + \alpha_2 V_1$ so the return per dollar is

$$r^* \frac{B_2}{V_1 - (V_2 - B_2)} = r^* \frac{1}{1 + \dfrac{V_1 - V_2}{B_2}}$$

If $V_1 < V_2$, the individual can obtain a perfectly safe return in excess of r^*. It follows

[6] Two firms, i and j are in the same risk class if $X_i(\theta) = \lambda X_j(\theta)$ for all θ. In the remainder of the discussion we shall assume, for convenience, that $\lambda = 1$.

[7] As usual, we assume no transactions costs and that there is no cash margin requirement on short sales. (See fn. 9.)

immediately that equilibrium in the capital market requires $V_1 = V_2$.[8]

Similar arguments can be used to show the following.

> If there are three or more firms in the same risk class, and the firms with the highest and lowest debt equity ratios have the same value, then the value of all other firms must be the same.

This is true whether individuals can borrow or can sell securities short. This result rules out the possibility of a U-shaped curve relating the value of the firm to the debt-equity ratio.[9]

III. *Bankruptcy*

Bankruptcy presents a problem for the usual proofs of the M-M theorem on two accounts: first, it means that the nominal rate of interest which the firm must pay on its bonds will increase as the number of bonds increases. (M-M have treated the case where it increases at exactly the same rate for all firms and individuals.) Second, if a firm goes bankrupt, it is no longer possible for an individual to replicate the exact patterns of returns, except if he can buy on margin, using the security as collateral; and if he defaults, he only forfeits the security and none of his other assets. To see this, consider the two alternative policies considered in Section I; in the one case, the firm issues no bonds (hence no chance of default) and in the other it issues $\hat{B}$ bonds. We have shown how the individual by buying stock on margin in the latter case can exactly replicate the returns in the former situation in those states where the firm does not go bankrupt if the value of the firm is the same in the two situations. But if the firm goes bankrupt in some state, θ', in the one case his return is zero, while in the other his return per dollar invested is

$$\frac{X(\theta')}{V}\left(1 + \frac{\hat{B}}{\hat{E}}\right) - r\frac{\hat{B}}{\hat{E}} < 0.$$

If, however, he can forfeit the security then his return will again be zero.

Of course, if the firm has a positive probability of going bankrupt, it will have to pay a higher nominal rate of interest. But if the individual is to use the security as collateral, he, too, will have to pay a higher nominal rate of interest. And indeed, it is clear that the two will be exactly the same, since the pattern of returns on the bonds in bankruptcy will be the same. Thus, we have shown that

> if a firm has a positive probability of going bankrupt, and an individual can borrow using those securities as collateral (so that if his return from the securities is less than his borrowings, he can forfeit the securities) the value of the firm is invariant to the debt-equity ratio.

It should be noted that the validity of this proposition does not require 100 percent margins. The required margin is only $\hat{B}/V$.

Individuals may, of course, not be able to make the limited liability arrangements or to obtain the level of margin required

[8] This proof has the advantage that no restrictions on the sign of $X(\theta)$ need be made. Although the case of $X(\theta) < 0$ is not very interesting from an economic point of view, some authors (e.g. [21]) have drawn attention to the difficulties which arise in the original M-M proof when $X(\theta) < 0$.

[9] Taxes and bankruptcy may alter this conclusion. Recently, Baumol and Malkeil [2] have argued that if there are costs of transactions, the levered company may have a higher value than the unlevered company. They argue that if, in order to undertake the arbitrage operations required by the M-M analysis, the individual had to borrow, the total value of transactions would be greater than if the company provided the desired leverage. If there are sizeable transactions costs, in order for the net income from the two firms to be the same, the levered company must have a higher value than the unlevered firm. Transactions costs cannot be adequately analyzed in terms of the two-period model that they (and we) use, but even in the context of a two-period model, it is not clear that their point is correct. If the individual has bonds in his portfolio, or if there are two companies, one with high leverage and one with low leverage, the individual can simply change his portfolio composition.

by the above analysis. Then, a firm by pursuing alternative debt-equity policies may be able to offer patterns of returns which the individual cannot obtain in any other manner (i.e. by purchasing shares in one or more other firms), and the value of the firm may consequently vary as the firm changes its debt-equity ratio. In the following subsections, we consider some special situations in which M-M results may still be valid, even though there is a finite probability of bankruptcy.

Risk Classes

If there are a large number of firms in the same risk class, then potentially they can all supply the same pattern of returns. If all firms maximize their value, then in market equilibrium all firms will have the same value.[10] Firms may have different debt-equity ratios and the same value for a number of reasons. For instance, assume that some individuals, for some reason or other, prefer a low debt equity-ratio, and some prefer a high debt-equity ratio. Then, some firms may have a high debt-equity ratio, some a low one. If one firm observes another firm in the same risk class with a different debt-equity ratio but a higher value, it will change its debt-equity ratio. Thus the observation that all firms in a given risk class have the same value but possibly different debt-equity ratios can be taken as evidence that firms are value maximizers and are in market equilibrium. It is not necessarily evidence that the arbitrage activities described by Modigliani and Miller have occurred, or that the value of the firm would be the same at some debt-equity ratio other than those actually observed.

Assume the market is in equilibrium, with $V=\rho EX$ for all members of the risk class. The securities sold by a firm are completely described by the risk class and the debt-equity ratio. A new, small firm is created, belonging to the same risk class, with mean return $\bar{X}$. If it chooses a debt-equity ratio used by other firms in the same risk class, the price of its shares must be the same as those of the other firms (since they are identical) so its value will be $\bar{X}\rho$. But, if it chooses some other debt-equity ratio, its value may be lower (if, for instance, there is a positive probability of bankruptcy).[11]

Mean-Variance Analysis and the Separation Theorem

In this subsection we consider the special case where all individuals evaluate alternative income patterns in terms of their mean and variance. For simplicty, let us assume that only the first firm issues enough bonds to go bankrupt. If all individuals agree on the probability distribution of returns for each firm, it can be shown that the

> . . . total market value of any stock in equilibrium is equal to the *capitalization* at the *risk-free interest rate* r^*, of the *certainty equivalent* . . . of its uncertain *aggregate dollar return;* . . . the difference . . . between the expected value of these returns and their certainty equivalent is *proportional* for each company to *its aggregate risk* represented by the *sum* of the *variance* of these returns and their total covariance with those of all other stocks, and the factor of proportionality is the *same* for *all* companies in the market. [11, pp. 26–27].

This implies that

$$(10)\quad E_i + B_i = \left\{\bar{X} - k\sum_{j=1}^{n} \mathcal{E}(X_i - \bar{X}_i)\,(X_j - \bar{X}_j)\right\} \Big/ r^* \qquad i = 2, \ldots, n$$

[10] Recall that we have assumed for expositional convenience that $X_j(\theta) = X_i(\theta)$ for all firms in the risk class.

[11] This also may occur with taxes if interest payments are tax deductible and if capital gains are treated preferentially. See [8].

(11) $E_1 =$

$$\left\{\bar{Z} - k \sum_{j=1}^{n} \mathcal{E}(Z - \bar{Z}) \ \ (X_j - \bar{X}_j)\right\} \Big/ r^*$$

(12) $B_1 =$

$$\left\{\bar{\hat{r}} B_1 - k \sum_{j=1}^{n} \mathcal{E}(\hat{r} - \bar{\hat{r}}) \ B_1(X_j - \bar{X}_j)\right\} \Big/ r^*$$

where $\mathcal{E}$ is the expections operator,

$$Z = max(X_1 - \hat{r}B_1, 0), \qquad \mathcal{E}Z = \bar{Z},$$
$$\mathcal{E}X_j = \bar{X}_j, \qquad \mathcal{E}\hat{r} = \bar{\hat{r}}.$$

and

$$k = r^* \left(\sum_i (X_i - \bar{X}_i) \Big/ \sum_i \sum_j \mathcal{E}(X_i - \bar{X}_i)(X_j - \bar{X}_j) \right).$$

Then adding B_1 and E_1, ((11) and (12)), we obtain

$$V_1 = E_1 + B_1 = \left\{\bar{X}_1 - k \sum_{j=1}^{n} \mathcal{E}(X_1 - \bar{X}_1) (X_j - \bar{X}_j)\right\} \Big/ r^*$$

independent of the debt-equity ratio.

The intuitive reason for this should be clear: it is well-known that if all individuals agree on the probability distribution of the risky assets, if there exists a safe asset, and if individuals evaluate income patterns in terms of mean and variance, the ratio in which different risky assets are purchased will be the same for all individuals, i.e. all the relevant market opportunities can be provided by the safe asset and a single mutual fund which (in market equilibrium) will contain all the risky assets, including the risky bonds. More generally, whenever the ratio in which different risky assets are purchased is the same for all individuals, then the M-M theorem will be true even with bankruptcy. For a complete discussion of the conditions under which the separation theorem obtains, see Cass and Stiglitz [4].

If, however, (a) all individuals do not agree on the probability distribution of $X_i(\theta)$ or, (b) the conditions under which the separation theorem is valid do not obtain, then the value of the firm will in general depend on the debt-equity ratio.[12]

Arrow-Debreu Securities

Arrow [1] and Debreu [5] have formulated a model of general equilibrium under uncertainty in which individuals can buy and sell promises to pay if a given state of the world occurs. See also Hirshleifer [10]. A stock market security and a bond can be viewed as a bundle of these Arrow-Debreu securities. If there is a sufficient number of different firms, equal to or greater than the number of states of na-

[12] For then, issuing a risky bond (a high debt-equity ratio) changes the relevant market opportunities available to the individual. For the M-M result to be valid, the debt-equity ratio can have no real effects on the economy. But it is easy to show that the assumptions (a) marginal utility of income in each state of nature is independent of the debt-equity ratio, and (b) the value of the firm is independent of the debt-equity ratio are in general inconsistent with the first order conditions for expected utility maximization being satisfied by all individuals (if bankruptcy may occur). To see this, observe that if an individual chooses his portfolio to maximize $\mathcal{E}U(Y(\theta))$, where $U''<0$, then a necessary and sufficient condition for the optimal allocation (assuming short sales are allowed) may be written $\mathcal{E}U'e_j = \mathcal{E}U'r^*$, or from (2), $[\underset{\mathcal{S}}{\mathcal{E}}U'(X_j - \hat{r}_j B_j)]/\mathcal{E}U'r^* = E_j$ where $\mathcal{S}$ is defined as the set of states of nature for which $X_j(\theta) \geq \hat{r}_j B_j$. Assume $U'(Y(\theta))$ is invariant for all θ and for all individuals to the j^{th} firm's debt-equity ratio. Then,

$$dE_j/dB_j = -\frac{\hat{r}_j}{r^*}\left(1 + \frac{d \ln \hat{r}_j}{d \ln B_j}\right)(\underset{\mathcal{S}}{\mathcal{E}}U'/\mathcal{E}U');$$

and if the value of the firm is to be unchanged, $dE_j/dB_j = -1$. But unless all individuals have identical utility functions *and* identical assessments of the probability of bankruptcy $(\underset{\mathcal{S}}{\mathcal{E}}U'/\mathcal{E}U')$ will differ for different individuals, so $dE_j/dB_j = -1$ only if marginal utilities in some states of nature change for some individuals.

It should be observed that when the actions of a firm can change the opportunity set, there is no reason that firms necessarily will maximize market value.

ture, then the market opportunities available to the individual (by purchasing or selling short different amounts of the market securities) are identical to those of a corresponding Arrow-Debreu market. If a promise to pay one dollar in state θ has a price $p^*(\theta)$,[13] then the value of the firm's equity is

$$E = \sum_{\mathcal{S}} (X(\theta) - rB)p^*(\theta).$$

If

$$r = \left[1 - \sum_{\mathcal{S}'} \frac{X(\theta)}{B} p^*(\theta)\right] \Big/ \sum_{\mathcal{S}} p^*(\theta)$$

where $\mathcal{S} \equiv \{\theta | X(\theta) \geq rB\}$, i.e. the states of nature in which the firm does not go bankrupt, and $\mathcal{S}' \equiv \{\theta | X(\theta) < rB\}$, then

$$E = \sum_{\theta} X(\theta)p^*(\theta) - B$$

i.e.

$$V = E + B = \sum_{\theta} X(\theta)p^*(\theta)$$

independent of the debt-equity ratio.

Three observations are in order: First, individuals do not need to agree on the probability of different states of nature occurring, i.e. they may disagree on the probability distribution of the returns to any firm.[14] Second, if there are fewer firms than states of nature, whether there are as many securities as states of nature is a function of the debt-equity ratio. If there are four states of nature and two firms, and if neither firm issues enough securities to go bankrupt, then there will only be three securities, but if one of the firms goes bankrupt, there will be four. Although the latter situation will be Pareto optimal (the marginal rate of substitution between consumption in any two states identical for all individuals), the value of the firm which goes bankrupt may be larger or smaller in the former situation than in the latter.[15]

Third, if we take literally the Arrow-Debreu definition of a state of nature, there undoubtedly will be more states of nature than firms. Yet, in some sense, most of these states are not very different from one another. For example, much of the variation in the return on stocks can be explained by the business cycle. If in any given business cycle state, the variance of the return were very small, and there were a small number of identifiable business cycle states, then the economy might look very much as if it were described by an Arrow-Debreu securities market.[16]

Bankruptcy and Perfect Capital Markets

The usual criterion for a perfectly competitive market is that the price of a commodity or factor an individual (or firm) buys or sells be independent of the amount bought or sold and be the same for all individuals in the economy. On this basis, it has been argued that the capital market is imperfectly competitive: (a) as a firm issues more bonds the rate of interest it pays may go up; (b) individuals may have to pay a higher interest rate than firms, and some firms higher than others; (c) lending rates may differ from borrowing rates. In this section, we have, however, considered perfectly competitive

[13] If there are no Arrow-Debreu securities on the market, $p^*(\theta)$ is the net cost to the individual of increasing his income in state θ by one dollar, i.e. by buying and selling short different securities. If there are more securities than states of nature, market equilibrium requires that the set of market prices generated by considering any subset of market securities which span the states of nature be independent of the particular subset chosen. For a more thorough discussion of these problems, see [4].

[14] They must, however, not assign zero probabilities to different states of nature occurring.

[15] In this situation we cannot assume that firms will necessarily maximize market value. (See fn. 12.)

[16] The point is that under these conditions the individual, by diversification of his portfolio, can essentially eliminate the variations in returns within a given business cycle state.

capital markets (with bankruptcy) in which all three of these would be true.[17] See also [22]. Thus the possibility of bankruptcy makes somewhat questionable the interpretation of much of this evidence of an imperfect capital market. The crucial fallacy lies in the implicit assumption that one firm's bond is identical to another firm's bond, and that bonds a firm issues when it has a low debt-equity ratio and those which it issues when it has a high debt-equity ratio are the same. But they are not. They give different patterns of returns. If there is any chance of default, a bond gives a variable return (i.e. is a risky asset). Just as there is no reason to expect butter and cheese, even though they are related commodities, to have the same price, so there is no reason to expect the nominal rate of interest where there is a low debt-equity ratio to be the same as when there is a high debt-equity ratio. Even the discrepancy between borrowing and lending rates does not imply imperfect capital markets, for when a person lends to the bank and the account is insured by FDIC, he can assume there is a zero probability of bankruptcy, but when the bank lends back to the same individual, it cannot make the same assumption.

REFERENCES

[1] K. J. Arrow, "The Role of Securities in the Optimal Allocation of Risk Bearing," *Rev. Econ. Stud.*, Apr. 1964, *31*, 91–96.

[2] W. Baumol and B. Malkiel, "The Firm's Optimal Debt-Equity Combination and the Cost of Capital," *Quart. J. Econ.*, Nov. 1967, *18*, 547–78.

[3] D. E. Brewer and J. B. Michaelson, "The Cost of Capital, Corporation Finance, and the Theory of Investment: Comment," *Amer. Econ. Rev.*, June 1965, *55*, 516–23.

[17] Transactions costs may also partly explain (b) and (c).

[4] D. Cass and J. E. Stiglitz, "The Structure of Preferences and Returns and Separability in Portfolio Allocation: A Contribution to the Pure Theory of Mutual Funds," *Cowles Foundation Discussion Paper*, May 1969.

[5] G. Debreu, *The Theory of Value*, New York 1959.

[6] P. Diamond, "The Role of a Stock Market in a General Equilibrium Model with Technological Uncertainty," *Amer. Econ. Rev.*, Sept. 1967, *57*, 759–76.

[7] D. Durand, "Cost of Capital, Corporation Finance, and the Theory of Investment: Comment," *Amer. Econ. Rev.*, Sept. 1959, *49*, 639–55.

[8] D. E. Farrar and L. L. Selwyn, "Taxes, Corporate Financial Policy, and Return to Investors," *Nat. Tax J.* Dec. 1967, *20*, 444–54.

[9] J. Hirshleifer, "Investment Decision under Uncertainty: Choice Theoretic Approaches," *Quart. J. Econ.*, Nov. 1965, *79*, 509–36.

[10] ———, "Investment Decision under Uncertainty: Applications of the State-Preference Approach," *Quart. J. Econ.*, May 1966, *80*, 237–77.

[11] J. Lintner, "The Valuation of Risk Assets and the Selection of Risky Investments in Stock Portfolios and Capital Budgets," *Rev. Econ. Statist.*, Feb. 1965, *47*, 13–37.

[12] H. Markowitz, *Portfolio Selection*, New York 1959.

[13] F. Modigliani and M. H. Miller, "The Cost of Capital, Corporation Finance, and the Theory of Investment," *Amer. Econ. Rev.*, June 1958, *48*, 261–97.

[14] ———, "Reply to Rose and Durand," *Amer. Econ. Rev.*, Sept. 1959, *49*, 665–69.

[15] ———, "Corporate Income Taxes and the Cost of Capital: A Correction," *Amer. Econ. Rev.*, June 1963, *53*, 433–43.

[16] ———, "Reply to D. E. Brewer and J. B. Michaelson," *Amer. Econ. Rev.*, June 1965, *55*, 524–27.

[17] J. MOSSIN, "Equilibrium in a Capital Asset Market," *Econometrica*, Oct. 1966, *34*, 768–83.

[18] G. PYE, "Investment Rules for Corporations," doctoral dissertation, M. I. T. 1963.

[19] A. A. ROBICHEK AND S. C. MYERS, "Problems in the Theory of Optimal Capital Structure," *J. Finance Quant. Anal.*, June 1966, *1*, 1–35.

[20] W. F. SHARPE, "Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk," *J. Finance*, Sept. 1964, *19*, 425–42.

[21] W. SHER, "The Cost of Capital and Corporation Finance Involving Risk." A paper presented at the winter meetings of the Econometric Society, Evanston, Illinois, Dec. 1968.

[22] G. STIGLER, "Imperfections in the Capital Market," *J. Polit. Econ.*, June 1967, *75*, 287–93.

Lags in the Effects of Monetary Policy: A Statistical Investigation

By J. ERNEST TANNER*

This paper is an attempt to study statistically the timing of the effects of monetary policy. In particular, it seeks to give an empirical answer to the question of when the money supply should be changed to effect predetermined changes in aggregate demand. This analysis requires an aggregate-dynamic model which takes explicit account of national income and the stock of money. The approach of this paper is to use augmented two-stage least squares to estimate a dynamic IS-LM model.

Recent empirical studies provide evidence that investment responds to changes in the level of interest rates, but with a distributed lag that may be too long for the monetary authorities to use monetary policy effectively for cyclical stabilization; see Thomas Mayer [23], J. H. Karaken and Robert Solow [2], and Harry Johnson and John Winder [19].[1] However, Donald Tucker [29] has recently argued that this assertion is not justified. Tucker points out that long distributed lags in the investment response can exist without causing comparable lags in the response of aggregate demand to changes in the money supply. Lags in the money demand function act to counter the investment lags. The argument is that, if the demand for money responds to interest rates with a distributed lag, a change in the money supply will temporarily force the interest rate beyond its new equilibrium level in order to equilibrate the market for money balances. This initial exaggerated interest rate response causes a correspondingly exaggerated effect on the components of aggregate demand and, hence, national income. Tucker's argument implies that despite long lags in the bulk of the investment response, sharp changes in the money supply may effect sharp changes in the rate of interest and, so long as the coefficient of the short-run response of investment to the interest rate is not zero, sharp short-run movements in aggregate demand.

This study builds on Tucker's analysis. In Section I, a general distributed lag form is developed and is used in the structural equations of the dynamic macromodel presented in Section II. Some of the estimation problems are discussed in Section III and the empirical estimates and their implications of the model for monetary policy are given in Section IV. Section V contains some concluding observations.

I. *Development of the Distributed Lag Function*

It has been argued that a major shortcoming of the popular geometric distributed lag function developed by L. M. Koyck [20] is that the coefficients of the

* The author is assistant professor of economics at the University of Western Ontario. He is deeply indebted to H. I. Grossman, M. B. Schupack, R. Sato, V. A. Bonomo, P. Cagan, J. Fried, G. Horwich and to the referee for useful suggestions. All errors are the author's.

[1] Following a different line of approach, Bonomo and Schotta [1] used spectral analysis in their study of the outside lag in the effects of monetary policy. They found that changes in the money supply tended to precede changes in income by only about two months. Unlike other studies, Bonomo and Schotta use real output (index of industrial production) rather than money national income as the series crossed against money. An explanation offered for the very short lag when compared to other studies is that prices tend to respond to monetary changes but only with a lag.

lagged values are required to diminish steadily. Instead of using the simple geometric lag distribution, this paper employs a generalized form of the negative binomial lag distribution developed by Solow [26].[2] The form of the lag distribution results from incorporating both adaptive expectations and partial adjustment mechanisms.[3]

If no costs were associated with rapid adjustments but costs were associated with not being in equilibrium positions, unconstrained optimal quantities of investment, consumption, and money balances would be demanded at all times. However, because there are costs in making asset transactions, attempted adjustments to newly established optimum quantities will be gradual.[4] The adjustment period will vary directly with the costs of adjustment and inversely with the costs associated with disequilibrium positions. During such periods of adjustment, constrained optimal quantities will be demanded. For expositional purposes, consider the asset X. The partial adjustment model specifies that the current constrained optimal quantity demanded X_t is related to the current unconstrained optimal quantity X^*_t by a coefficient of adjustment $(1-\lambda)$,

$$(1) \quad X_t = X_{t-1} + (1-\lambda)(X_t^* - X_{t-1}).$$

The current unconstrained optimal quantity X^*_t is assumed in turn to depend upon an expected return $Z_{t+1}{}^E$ which is the value of Z currently expected to prevail in subsequent periods:

$$(2) \qquad X_t^* = a + bZ_{t+1}^E.$$

We relate $Z_{t+1}{}^E$ to observable variables by specifying the expectation generating equation:

$$(3) \quad Z_{t+1}^E = Z_t^E + (1-\beta)(Z_t - Z_t^E).$$

The adaptive expectations model given by this equation says that expectations about future returns are revised in proportion to the error associated with previous levels of expectations. John Muth [24] has shown that if the process generating the measured value of Z is such that the change in the measured value of Z is a first-order moving average of random deviates, then the expectation generating function described by equation (3) is an optimal forecast of the measured value of Z. In fact, equation (3) not only gives the best forecast of expected return for the period immediately ahead, but it also gives the best forecast for any future time period. This is so because the forecast gives an estimate of the permanent component of Z.

The combination of these three equations yields the reduced form of a general negative binomial lag distribution:[5]

[2] The lag distribution used in this paper would be identical to the negative binomial lag distribution Solow [26] developed and used in his joint study with John Kareken for the Commission on Money and Credit [2], if the coefficient of adjustment were equal to the coefficient of expectations.

[3] The theoretical underpinnings of the lag distribution follows the work of Feige [12].

[4] Consider the case for lags in the adjustment of money balances. A certain amount of inertia is introduced into portfolio composition due to the fact that "Every reconsideration of the portfolio involves the investor in the expenditure of time and effort as well as of money . . ." Tobin [28, p. 67]. The frequency of portfolio review and adjustment will, therefore, depend upon such factors as: 1) the individual, 2) the size of the portfolio, 3) costs of engaging in transactions, 4) opportunity costs of the investor's time, and 5) the magnitude of foregone income from the portfolio's current composition. The upshot of all this is that many portfolios will not be changed to take advantage of minor transitory fluctuations in the explanatory variables.

[5] It is obvious from equation (4) that the partial adjustment model and the adaptive expectations model are special cases. The partial adjustment model is the special case where expectations are static, i.e., where $\beta=0$. Similarly, the adaptive expectations model occurs when $\lambda=0$. It has been pointed out by Waud [30] that estimating the special case models (i.e., the adaptive expectations or the partial adjustment model), when the true model is in fact a combination of both as given

$$(4)\quad \begin{aligned} X_t = {} & (1-\lambda)(1-\beta)(a+bZ_t) \\ & + (\lambda+\beta)X_{t-1} - \lambda\beta X_{t-2}. \end{aligned}$$

The form of the lag distribution given by this equation is used in each of the estimated behavioral equations of the model.

II. *The Aggregate Dynamic Model*

In order to give empirical content to Tucker's offsetting lag hypothesis, we require a macro-model which allows for dynamic adjustments in both the commodity and monetary sectors. An analysis of the exact effects of money supply changes upon aggregate expenditures and output requires a very large dynamic econometric model. Nevertheless the essential features of the offsetting lag hypothesis can be captured by a very simple dynamic form of the familiar IS-LM model. Therefore, the framework will be kept as simple as possible within the bounds of retaining the basic idea of Tucker's offsetting lag hypothesis. Consequently only two dynamic structural equations are employed, a private expenditure (the sum of consumption plus investment demands) function and a money demand function. Other variables such as the government spending, taxes, and foreign trade balance variables are treated as exogenous. In addition, labor market adjustments and capital utilization changes are ignored. The focus of the model is solely on the question of the timing relationship between changes in the money supply and changes in aggregate demand.[6] And so, the effects of money supply changes on income are not broken down into changes in real output and changes in prices.

in equation (4), results in a serious bias in the least squares estimates of the regression coefficients. Furthermore, the misspecification leads to an upward bias in the estimated lag and to a substantial increase in the size of the estimated standard errors relative to the estimated regression coefficients.

[6] The lag in the effect of monetary policy considered in this paper is only that part of the lag which occurs *after* the money supply is changed. Because Bryan [3] found that the dynamic responses of the banking system to policy changes of the Federal Reserve System are relatively short, the part of the lag occurring between policy actions of the central bank and changes in the money supply are not analyzed here.

The model uses a dynamic form of the static IS-LM framework to determine the timing relationship between changes in the money supply and changes in aggregate demand. This is accomplished by incorporating the distributed lag relationship developed above into the aggregate structural relations of the existing textbook type of macro-model. In addition to a private expenditure function and a money demand function, the model postulates two equilibrating conditions. These four relationships determine the endogenous variables: the private expenditure, money demand, rate of interest, and national income. Other equations are added to complete the model and to specify the exogenous variables.

$$(5)\quad \begin{aligned} X_t = {} & \\ & (1-\lambda)(1-\beta)\cdot(A_1 + B_1[Y-T]_t + D_1r_t) \\ & + (\lambda+\beta)X_{t-1} - \lambda\beta X_{t-2} \end{aligned}$$

$$(6)\quad \begin{aligned} M_t^d = {} & \\ & (1-\gamma)(1-\phi)\cdot(A_2 + B_2Y_t + D_2r_t) \\ & + (\gamma+\phi)M_{t-1} - \gamma\phi M_{t-2} \end{aligned}$$

$$(7)\quad M_t^d = M_t^s$$

$$(8)\quad M_t^s \equiv \overline{M}_t^s$$

$$(9)\quad Y_t = X_t + G_t + E_t$$

$$(10)\quad G_t \equiv \overline{G}_t$$

$$(11)\quad T_t \equiv \overline{T}_t$$

$$(12)\quad E_t \equiv \overline{E}_t$$

The variables are as follows:

X: aggregate private commodity expenditures,
Y: national income,
r: interest rate,
M^d: money demand,
M^s: money supply,

G: government expenditures,
T: government taxes,
E: net exports over imports,
t: time subscript,

ϕ and γ are distributed lag parameters in the money demand function and λ and β are the lag parameters of the commodity market where $0 \leq \lambda, \beta, \gamma, \phi < 1$. All other symbols are constants.

To estimate the structural relationships of the model, we must decide upon the statistical series to best measure the aggregate variables. Consider the identity

$$GNP = C + I + G + E \tag{13}$$

with consumption, gross investment, government purchases, and net exports (exports less imports) as four components.

Because the income variable appearing in the money demand function (equation (6)) is a proxy for transactions, this definition of the income variable appears to be the appropriate definition. However, it seems more reasonable to assume that the income variable used to explain aggregate private expenditures be net national product less government taxes net of transfers:

$$NNP - T = C + I^N + G - T + E \tag{14}$$

where I^N is net investment and T is government taxes net of transfer payments. This definition of income differs from the widely used disposable income variable by including undistributed corporate profits. Whether we should include undistributed corporate profits is an open question. Although it seems best to leave it out of the income variable explaining the consumption component of aggregate private expenditures in equation (5), it appears to strongly influence the investment component.[7] For this reason, we include undistributed corporate profits in our income series which is used to explain aggregate private expenditures. However, the income series we actually use is $C+I+G-T+E$ rather than $C+I^N+G-T+E$. We use gross investment rather than net investment because the data on capital depreciation are poor and because both series probably move rather closely over the sample period, 1947 through 1967.

The answer to the question as to which series should be used to represent the stock of money is not clear. In fact, recent research using different monetary series has not arrived at a theoretically or empirically best series. However, since we seek appropriate monetary policy to stabilize aggregate economic activity, we shall use the monetary series consisting of only demand deposits and currency. These assets represent not only the means of payment but they also bear the first effects of the Federal Reserve System's policy. For these reasons, the narrow definition of money is used.

The series used as a proxy for the theoretical interest rate variable is also arbitrary. However, it should be a series with considerable variation which moves in substantial uniformity with the structure of interest rates. The particular series chosen is the bankers' acceptances prime ninety day rate.[8]

In estimating the model, we assume that the authorities are able to control the money supply and it is determined exogenously.[9] We further assume that the

[7] Numerous studies have found that internally generated funds or profits have an observable effect on investment for given levels of output, capital stock, and interst rate. References to and major conclusions of empirical investment studies are made in Edwin Kuh [21].

[8] Because Keynesian economics links a long term interest rate to investment and income through the demand for and supply of money, some researchers favor a long-term interest rate on bonds to be used in macromodels. However, in the present case, since an expectations model is used, the expected short rate at a point in time can be thought of as a current long term interest rate.

[9] The money supply is not truly an exogenous variable and, to this extent, our regression estimates may not be consistent. However, since our purpose is not to build a model determining the supply of money, such an assumption is convenient if not entirely correct.

government and foreign sectors are exogenous. Thus, our model determines the rate of interest, the level of national income, money demand and aggregate private expenditures as endogenous variables.

III. *Estimation Problems and Procedure*

Previous empirical studies dealing with the lag in the effect of monetary policy have employed single equation models. Because such relationships ignored all other sectors of the economy, they suffer from single equation biases. In contrast, the present model contains two sectors and the estimation procedure uses two-stage least squares in order to take explicit account of the interrelation between the commodity and money markets. In the two-stage least squares procedure, the values of the interest rate and income (endogenous variables used as explanatory variables in the second stage) used for estimating the two behavioral relationships are the predicted values obtained by regressing these variables on the model's predetermined variables.[10]

The variables which can be legitimately called predetermined variables differ depending upon the assumptions made. If no autocorrelation is assumed in the error term U_t of the commodity demand equation,

$$(5') \quad X_t = (1-\lambda)(1-\beta)\,(A_1 + B_1\,[Y-T]_t + D_1 r_t) + (\lambda+\beta)X_{t-1} - \lambda\beta X_{t-2} + U_t$$

we will obtain consistent estimates in the second stage if we treat X_{t-1} and X_{t-2} in the first stage as predetermined.[11] However it is very unlikely for U_t to be serially uncorrelated[12] and because no statistic exists which can be used uncritically to test for the presence of serial correlation in the residuals of relationships containing lagged endogenous variables,[13] the model is estimated using the method of augmented least squares which yields consistent estimates of the parameters without placing the restriction of randomness on the distribution of U_t.[14]

[10] In the method of augmented two-stage least squares, the first stage predetermined variables used for obtaining the predicted values of the endogenous explanatory variables include only the model's exogenous variables M_t, M_{t-1}, M_{t-2}, E_t, and G-T_t. We do not treat X_{t-1} and X_{t-2} as predetermined because we make no assumption about randomness in the error term of the commodity demand equation. See fnn. 11 and 12.

[11] In equation (5′), observe that current aggregate private demand values are influenced by current disturbances so that lagged private demand values are influenced by lagged disturbances. If U_t is not correlated with its own past values, then the covariance of the elements of the lagged expenditure values and the current values of the error term are a sum of elements each of which involves the covariance of U_t with either past exogenous variables or with past disturbances. All such terms are zero in the probability limit. Hence, if U_t is not autocorrelated, consistent estimates will be obtained by simply treating X_{t-1} and X_{t-2} as exogenous in the estimation procedure.

[12] Consider a simplified form of equation (5′) where the unconstrained optimal value of X depends upon only the expected values of Y plus an error term.

$$(5a) \quad X_t^* = a + bY_{t+1}^E + V_t.$$

Let $(1-\beta)$ and $(1-\lambda)$ be the coefficient of expectations and the coefficient of adjustment respectively. In terms of observable variables, (5a) becomes

$$(5b) \quad X_t = (1-\lambda)(1-\beta)(a+bY_t) + (\lambda+\beta)X_{t-1} - \lambda\beta X_{t-2} + (1-\lambda)[V_t + \beta V_{t-1}].$$

For the error term of (5b) not to be serially correlated, then the error term of (5a) must be autocorrelated with the form

$$(5c) \quad V_t = \beta V_{t-1} + W_t$$

where W_t is a truly random disturbance. If the error term does not follow this first-order Markov scheme, then the application of ordinary two-stage least squares, where X_{t-1} and X_{t-2} are treated as predetermined in the first stage, will not yield consistent estimates of the parameters.

[13] The Durbin-Watson statistic does not warn us of possible autocorrelation in the commodity demand equation. The statistic is biased towards 2 (the value it should have if no serial correlation is in fact present) when the estimated relationships contain lagged endogenous variables. See Nerlove and Wallis [25].

[14] See Feldstein [13] for a proof of the consistency of augmented least squares.

Suppose we wish to estimate the following correctly specified relationships:

$$(5'a) \quad X_t = B_0 + B_1Y_t + B_2X_{t-1} + V_t$$

and

IV. *The Empirical Results*

1. *The Estimates*

Using quarterly data from the United States for the period 1947 through 1967,[15] the augmented two-stage least squares estimation procedure yielded the following structural equations:

$$
(15)\quad
\begin{aligned}
X_t = &- \underset{}{11.661} - \underset{(4.01)}{8.270\hat{r}_t} \\
&+ \underset{(4.13)}{0.315\widehat{[Y - T]}_t} + \underset{(13.66)}{1.227X_{t-1}} \\
&- \underset{(5.63)}{0.510X_{t-2}} + \underset{(6.16)}{0.493\hat{U}_t} \\
&- \underset{(4.39)}{0.300\hat{U}_{t-1}} \qquad R = .999 \quad F = 10681.756
\end{aligned}
$$

$$
(16)\quad
\begin{aligned}
M_t = &\ 19.411 + \underset{(10.98)}{0.042\hat{Y}_t} - \underset{(8.84)}{1.670\hat{r}_t} \\
&+ \underset{(19.74)}{1.311M_{t-1}} - \underset{(7.76)}{0.491M_{t-2}} \\
&\qquad R = 0.999 \quad F = 29154.261
\end{aligned}
$$

The values of the Student's t-statistic (the absolute value of the ratio of the regression coefficient to its standard error) are given directly below the regression coefficients.

All the estimated coefficients are of the desired sign (i.e., positive for income, negative for the rate of interest and alternating for lagged variables) and significantly different from zero at the 5 percent level. The R and F values are exceedingly high implying the overall fit of the model to the data is good and the explained variance of the dependent variables are far greater than might be accounted for by chance. But, as is well-known, the lagged variables in each equation contribute to these high R and F statistics.

2. *Implied Time Paths of the Money Supply and Interest Rate*

In order to use these estimated structural equations to give an answer to the question of the lag in the effects of monetary policy, we combine the equations of the model (eliminating income and the interest rate) to derive an explicit expression relating private expenditures to the money supply.

$$
(17)\quad
\begin{aligned}
M_t = &\ A + BX_t - CX_{t-1} + DX_{t-2} \\
&+ (\gamma + \phi)M_{t-1} - \gamma\phi M_{t-2}
\end{aligned}
$$

where

$$
A = (1-\gamma)(1-\phi)\left[A_2 + B_2G_t + B_2E_t - D_2\left(\frac{A_1}{D_1} + B_1\frac{E_t + (G-T)_t}{D_1}\right)\right]
$$

$$
B = \frac{D_2(1-\phi)(1-\gamma)}{D_1(1-\beta)(1-\lambda)} \cdot \left[1 - (1-\beta)(1-\lambda)\left(B_1 - \frac{D_1B_2}{D_2}\right)\right]
$$

$$
C = \frac{D_2(1-\gamma)(1-\phi)}{D_1(1-\beta)(1-\lambda)}(\beta + \lambda)
$$

$$V_t = \sum_{j=1}^{n} p_jV_{t-j} + W_t$$

where W_t is serially independent, Y_t is exogenously given, and observations on V_t are available. The coefficients of (5′a) can be consistently estimated by ordinary least squares.

In the present situation, the time-series V_t is a series of unobserved disturbance terms. In the method of augmented least squares, we first obtain a consistent series for V_t:

$$(5'b)\qquad \hat{V}_t = X_t - b_0 - b_1Y_t - b_2X_{t-1}$$

where the b's are consistent estimates of the B's. The instrumental variable method (replace X_{t-1} by the predicted value obtained by regressing X_{t-1} on Y_t and the instrumental variable Y_{t-1} when estimating (5′b) provides consistent coefficient estimates of the b's (see Liviatan [22]) allowing the residual series $\hat{V}_t$ to be used in place of V_t in (5′a). Consistent estimates of the distributed lag relationship are, then, obtained by applying ordinary least squares to the transformed version of (5′a).

[15] The exact data used and method of estimation are available from the author upon request.

$$D = \frac{D_2(1-\gamma)(1-\phi)}{D_1(1-\beta)(1-\lambda)} \lambda\beta$$

The explicit solution of equation (17) requires that we specify an initial value for M and a complete time path for X. We seek the required behavior of the money supply over time on its path to the new equilibrium to effect a once and for all desired change in aggregate expenditures. The special solution in which the money supply is at equilibrium before income is to change and is allowed to reach its new equilibrium after effecting the change in income is not very useful for illuminating an answer to our basic question. It is the behavior of the required changes in the money supply over time which is useful to us, not the new static equilibrium level. The question we want to answer is this: Given that in some future time period we desire income to change from what it would have been in the absence of discretionary monetary policy, how would the path of the money supply differ over time in order to effect this desired income change? In other terms, how will the money supply be required to change over time from what otherwise would have prevailed to effect the desired predetermined change in income?

In analyzing the dynamic behavior of the money supply, we are more interested in the path it follows to equilibrium than in the absolute magnitude of individual changes. Therefore, we shall express the change in the money supply from what it otherwise would have followed in any time period as a proportion of the equilibrium change in the money supply:

$$\frac{M_t - M_0}{M_E - M_0}$$

where M_t is the required money supply in period t, M_E is the equilibrium money supply, and M_0 is the path the money supply would otherwise have followed in the absence of discretionary policy. Alternatively, we assume that we desire a once and for all increase in income of ΔX occurring at time $t=1$. This higher level of income will be maintained indefinitely into the future. By setting

$$\begin{aligned} X_t - X_0 &= X_{t-1} - X_0 \\ &= X_{t-2} - X_0 = \Delta X \text{ and} \\ M_t - M_0 &= M_{t-1} - M_0 \\ &= M_{t-2} - M_0 = \Delta M, \end{aligned}$$

we calculate the equilibrium change in the money supply to be[16]

$$(18) \quad M_E - M_0 = \Delta M_E = \frac{B-C+D}{(1-\gamma)(1-\phi)} \Delta X.$$

As a proportion of the equilibrium change, the money supply in the first period is

$$(19) \qquad \frac{M_1 - M_0}{M_E - M_0} = \frac{B(1-\gamma)(1-\phi)}{B - C + D}$$

Similarly, the money supply in period two, . . . in period n is:[17]

[16] By change in the money supply in any time period, we mean the difference between the required money supply to effect the desired time path of income and what the money supply would otherwise have been. Therefore, the equilibrium change in the money supply refers to the difference between the equilibrium required stock of money and what would have prevailed in the absence of discretionary policy and does not refer to a continuing change in the money supply.

[17] This paper deals solely with the suitability of monetary policy for cyclical stabilizing requirements. Thus the dynamic properties of the model are applicable only to short-run analysis. However, for those interested in the model's long-run growth properties, we can use equation (17) to obtain the money/income steady state solution. Assume that the money supply and income grow at rate α, then we obtain

$$(17') \quad \frac{M_t}{X_t} = \frac{1}{X_t}\left[\frac{A}{1-(\gamma+\phi)(1+\alpha)+\gamma\phi(1+\alpha)^2}\right] + \frac{[B - C(1+\alpha) + D(1+\alpha)^2]}{1-(\gamma+\phi)(1+\alpha)+\gamma\phi(1+\alpha)^2}$$

If the exogenous balances (i.e., the government and the international balance) are constant, the first term goes to zero when $\alpha>0$, and the steady state solution is simply

$$\frac{M_2-M_0}{M_E-M_0}=(\gamma+\phi)\,\frac{M_1-M_0}{M_E-M_0}+\frac{(B-C)(1-\gamma)(1-\phi)}{B-C+D} \tag{20}$$

$$\vdots$$

$$\frac{M_n-M_0}{M_E-M_0}=1+(\gamma+\phi)\left[\frac{M_{n-1}-M_0}{M_E-M_0}-1\right]-(\gamma\phi)\left[\frac{M_{n-2}-M_0}{M_E-M_0}-1\right] \tag{21}$$

for $n=3, 4, 5, \ldots$.

Substitute the estimated values of the parameters[18,19]

$B_1 =$	1.113	$B_2 =$	0.233
$D_1 =$	-29.223	$D_2 =$	-9.277
$\lambda+\beta =$	1.227	$\gamma+\phi =$	1.311
$\lambda\beta =$	0.510	$\gamma\phi =$	0.491

into equations (19) through (21). This gives the time path to equilibrium of the money supply which is required to effect the permanent desired increase in income in period one. The time path of the money supply is plotted in Figure 1, Part A.

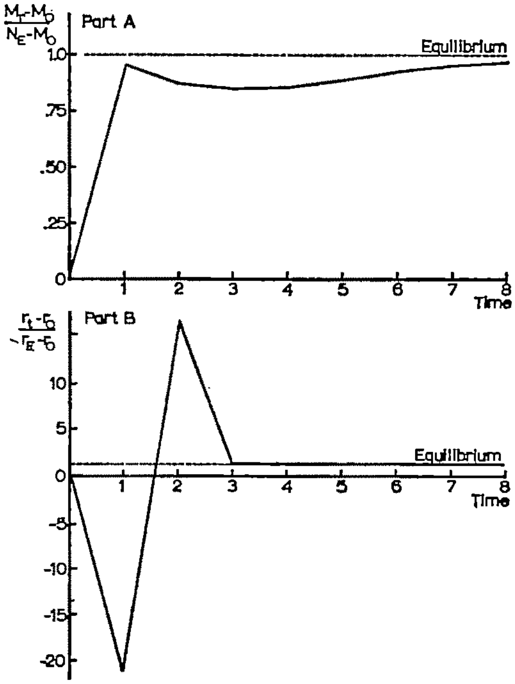


FIGURE 1

As the figure shows, the implied lag in monetary policy is extremely short. In fact, the impact effect of a change in the money supply is about equal to the long-run effect. The initial change in the money supply is over 95 percent of the required long-run equilibrium change. In the second period, the money supply is required to fall below equilibrium and it continues falling until period three when it is about 80 percent of the long-run value. In later periods, the money supply approaches equilibrium asymptotically.

The mechanism through which this behavior of the money supply will effect a once and for all permanent change in aggregate demand is the interest rate. The striking thing about the time path of the interest rate (diagrammed as Part B, Figure 1) is its behavior in the first two periods. Its movement mirrors neither the time path of the money supply nor the time path of aggregate demand. Because

---

$$\frac{B-C(1+\alpha)+D(1+\alpha)^2}{1-(\gamma+\phi)(1+\alpha)+\gamma\phi(1+\alpha)^2}.$$

Employing the empirical estimates and making the assumption of a 5 percent growth rate, the steady state money-income ratio is about 0.21. On the other hand, if the exogenous balances grow at the rate α, the steady state solution is as given in equation (17′).

[18] This estimate of the marginal propensity to spend ($B_1=1.113$) implies an infinite multiplier. In actual fact, it is unlikely that this would cause the system to be unstable in the sense of a true growth model. In the case of the type of model used in this paper, making imports and taxes endogenous would probably make the model stable. In the paper by Chow [7] where taxes are treated as endogenous, the tax rate for the period from 1943 to 1963 is estimated to be 0.21. This tax rate would imply a multiplier of 8.28.

$$\left[\frac{1}{1-1.113\ (1-0.21)}\right].$$

[19] The point estimates of the lag coefficients give imaginary solutions to λ, β, γ, and ϕ. However, the confidence intervals are such that they contain real solutions for the parameters.

of the long distributed lag in the money demand function, the interest rate violently overshoots its new equilibrium in period one. In the second period, the interest rate reverses direction with sufficient briskness to cause it again to overshoot equilibrium.[20] However, after the second period's reaction, the interest rate quickly converges to equilibrium.[21] It is this vigorous overshooting in the interest rate which causes the sharp, but permanent, change in aggregate demand. The changes in the interest rate affect the new level of aggregate demand through their effects on the investment component of aggregate private demand. Because investment responds with a distributed lag, and not a discrete lag, investment does depend (at least partially) upon the current rate of interest and, thus, sufficiently large changes in the interest rate are able to effect any predetermined change in investment.[22] However, the violent swings in the interest rate may not be practical for the monetary authorities. To those who feel the authorities should provide a stable and predictable financial environment as given by a relatively constant interest rate, an initial downward movement of 83 basis points of the interest rate to effect an immediate $10 billions increase in aggregate demand followed one quarter later by an even larger upward movement of the interest rate may be unacceptable. Nevertheless, if the full effects of a change in the money supply are to be felt almost immediately, the model requires these extraordinary swings in the interest rate.

[20] The same kind of interest rate behavior has recently been found by Cagan and Gandolfi [6]. They offer two reasons why the initial decline in the interest rate resulting from an increase in the money supply will reverse itself and be carried beyond the new higher equilibrium level. They argue that: (1) because of a lagged response of income to money supply changes, income must grow faster than money to make up for the initial lag and, thus, produce overshooting; and (2) an increase in the growth rate of the money supply will lead to a higher rate of price change which will lead to higher nominal interest rates to compensate for the depreciation in the real value of fixed-dollar assets.

[21] It should be noted that the equilibrium relationship between changes in the interest rate and changes in the money supply which is implied by the regression coefficients is positive. Although there is a negative association between concurrent changes in the money supply and interest rate, this concurrent effect is soon offset and becomes a positive relationship in equilibrium. The same phenomenon has also been found by Cagan [5], Cagan and Gandolfi [6], Gibson [15] and Gibson and Kaufman [16]. Gibson and Kaufman write "... the negative influence of money (on interest rates) fades away after only three months ... (and this explains the) ... surprising statements of some economists that high interest rates reflect easy monetary policy and low interest rates tight policy ..." [16, p. 478].

[22] This analysis assumes that the magnitude of the effects on investment are proportional to the absolute changes in the interest rate regardless of the absolute size of these changes. In other words, we assume that if a k percent decrease in the interest rate immediately raises aggregate demand by $$h$ billions, than a 100 k percent decrease in the interest rate will immediately raise aggregate demand by $100 h billions. To the extent that this assumption is false, reasonable changes in the interest rate may not be able to induce the desired short-run movements in aggregate demand. However, this does not require that monetary policy works slowly. Rather, it implies only that the magnitude of the short-run effects of monetary policy may be limited.

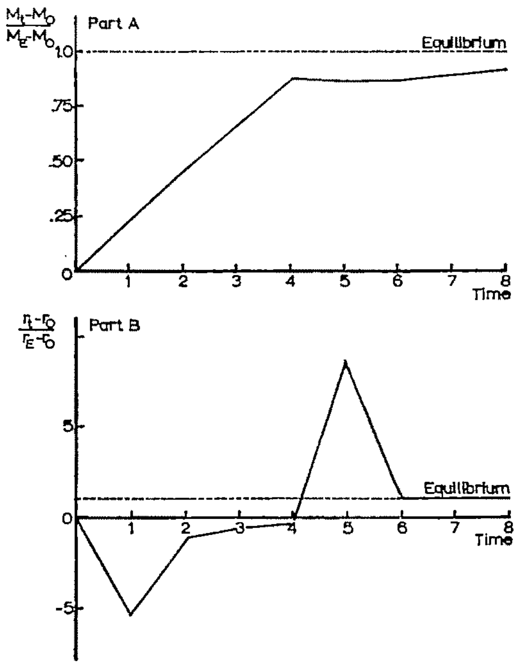


FIGURE 2

TABLE 1—REQUIRED INTEREST RATE TIME PATHS AS EXPRESSED IN BASIS POINTS FROM INITIAL VALUE FOR A $10 BILLIONS INCREASE IN AGGREGATE DEMAND EFFECTED THROUGH MONETARY POLICY

Effected in one Quarter	Effected in one Year ($2.5 billions per quarter)
$r_0=0.0$	$r_0=0.0$
$r_1=-82.8$	$r_1=-20.7$
$r_2=+65.5$	$r_2=-4.3$
$r_3=+3.87$	$r_3=-3.3$
$r_4=+3.87$	$r_4=-2.4$
$r_5=+3.87$	$r_5=+34.7$
$r_6=+3.87=r_B=$	$r_6=+3.87$

However, if the desired change in aggregate demand is allowed to occur over a year, rather than one quarter, then the quarter to quarter changes in the money supply and interest rate are no more than, and frequently less than, a quarter of the required changes when income must make the permanent change in three months. The time paths to equilibrium of the money supply and interest rate for four equal increments in income occurring in four successive three months' periods are diagrammed in Figure 2. As this figure shows, the time paths of the money supply and interest rate are much smoother and, therefore, more acceptable than the time paths diagrammed in Figure 1. For comparison, the actual time paths of the interest rate required to effect a permanent increase in aggregate demand by $10 billions for the two policies are given in Table 1.

V. *Summary and Concluding Observations*

This paper has attempted to provide empirical evidence on the timing of the effects of monetary policy. This was done by estimating a distributed lag dynamic macro-model which also allows the tracing out of the time path of the money supply required to effect a permanent change in national income. The macro-model used incorporated a general negative binomial lag relationship into the behavioral equations of the existing static Keynesian national income model. By combining augmented and two-stage least squares in estimating the structural relationships of the model, we attempted to eliminate many of the deficiencies of past research concerned with the timing of the effects of monetary policy. The employment of a multiple equation model and the estimation procedure used not only gives consistent estimates of each parameter but also takes account of the offsetting lags in the economy.

Recent empirical studies of the lag placed much emphasis on the lag in the investment function. The attention focused on this lag was based upon the belief that the lag in the effect of monetary policy was at least as long as the investment lag. However, as Tucker points out, substantial lags in the money demand function can partially or wholely negate the investment lag. Indeed, this paper, which also finds rather long lags in the investment function, finds that the bulk of the effects on aggregate demand of monetary policy can occur within three to six months of the change in the money supply. This conclusion is in contrast to earlier studies which found the short-run effects to be small relative to the long-run effects; so small in fact that active anticyclical monetary policy, with its heavy emphasis on short-run adjustments, is likely to be ineffective in achieving its objectives. The findings of this paper cast serious doubt on the conclusion that anticyclical monetary policy must be abandoned, entrusting the objective of short-run stability to the other instruments of economic policy such as fiscal policy or direct controls of one kind or another. In fact, the evidence suggests that the impact effect of a change in the money supply is of about the same magnitude as the long-run effect. To the extent that this conclusion is true, monetary policy is ideal

for and consistent with short-run stabilization requirements.

However, some qualifications on these conclusions must be made. First, as it has already been pointed out, although monetary policy probably works quickly, it may not be able to induce short-run changes in income of very large magnitudes. To the extent that the size of the effects may be limited in the short run, other cyclical stabilization policies, in addition to monetary policy, are required. The second limitation is the point that no adjustment was made for the fact that many specifications were tried before settling on this one.[23] As is well-known, this procedure induces an upward bias in the significance tests in comparison to the significance levels relevant under traditional assumptions. For simplicity, this study has not taken explicit account of either the foreign or government sector. Furthermore, it is incapable of indicating either how the money supply is changed or the exact channels through which monetary policy works. Because of these limitations, the present model is too simple to be used for exact policy decisions. However, the model does indicate that monetary policy does appear to work quickly enough for cyclical stabilization requirements and there should be little hesitation in applying monetary policy to this end.

[23] The results of the model are very sensitive to the definition of income in each structural equation. For example, the use of disposable income in the commodity demand equation caused the sign on the interest rate to become positive. In addition, if $\hat{U}_{t-1}$ is omitted, the interest rate variable becomes insignificant. However, this latter result does not occur if the time period is shortened to 1952 through 1966.

REFERENCES

1. V. A. Bonomo and C. Schotta, "Some Aspects of the Outside Lag in Effect of Monetary Policy: A Preliminary Report," *Proc. Amer. Statist. Ass.*, Aug. 1968, 482–90.
2. E. C. Brown, R. M. Solow, A. Ando, and J. H. Kareken, "Lags in Fiscal and Monetary Policy," in *Stabilization Policies*, for the Commission on Money and Credit. Englewood Cliffs 1963.
3. W. R. Bryan, "Bank Adjustments to Monetary Policy: Alternative Estimates of the Lag," *Amer. Econ. Rev.*, Sept. 1967, *57*, 855–64.
4. P. Cagan, "The Monetary Dynamics of Hyperinflations," in M. Friedman, ed., *Studies in the Quantity Theory of Money*. Chicago 1956.
5. ———, "The Channels of Monetary Effects on Interest Rates," Nov. 1966. Mimeo.
6. ———, and A. Gandolfi, "The Lag in Monetary Policy as Implied by the Time Pattern of Monetary Effects on Interest Rates," *Amer. Econ. Rev. Proc.*, May 1969, *59*, 277–84.
7. G. C. Chow, "Multiplier, Accelerator, and Liquidity Preference in the Determination of National Income in the United States," *Rev. Econ. Statist.*, Feb. 1967, *49*, 1–15.
8. C. Christ, "A Simple Macroeconomic Model with a Government Budget Constraint," *J. Polit. Econ.*, Jan–Feb. 1968, *76*, 53–67.
9. J. M. Culbertson, "Friedman on the Lag in Effect of Monetary Policy," *J. Polit. Econ.*, Dec. 1960, *68*, 617–21.
10. ———, "The Lag Effect of Monetary Policy: Reply," *J. Polit. Econ.*, Oct. 1961, *69*, 467–77.
11. F. DeLeeuw, "The Demand for Money—Speed of Adjustment, Interest Rates, and Wealth," Staff Economic Study, Board of Governors, Federal Reserve System, Oct. 1965. Reprinted in G. Horwich, ed., *Monetary Process and Policy: A Symposium*. Homewood 1966.
12. E. L. Feige, "Expectations and Adjustments in the Monetary Sector," *Amer. Econ. Rev. Proc.*, May 1967, *57*, 462–73.
13. M. S. Feldstein, "The Consistency of Augmented Least Squares," Harvard Univ., Nov. 1967, Mimeo.
14. M. Friedman, "The Lag in Effect of Monetary Policy," *J. Polit. Econ.*, Oct. 1961, *69*, 447–66.

15. W. Gibson, "Effects of Money on Interest Rates," Federal Reserve Bank of Chicago, Jan. 1968. Mimeo.
16. ———, and G. Kaufman, "The Sensitivity of Interest Rates to Changes in Money and Income," *J. Polit. Econ.*, May-June 1968, *76*, 472–78.
17. Z. Griliches, "A Note on the Serial Correlation Bias in Estimates of Distributed Lags," *Econometrica*, Jan. 1961, *29*, 65–73.
18. ———, "Distributed Lags: A Survey," *Econometrica*, Jan. 1967, *35*, 16–49.
19. H. G. Johnson, and J. W. L. Winder, *Lags in the Effect of Monetary Policy in Canada*, (working paper prepared for the Royal Commission on Banking and Finance) 1963.
20. L. M. Koyck, *Distributed Lags and Investment Analysis*. Amsterdam 1954.
21. E. Kuh, "Theory and Institutions in the Study of Investment Behavior," *Amer. Econ. Rev.*, May 1963, *53*, 260–68.
22. N. Liviatan, "Consistent Estimation of Distributed Lags" *Int. Econ. Rev.*, Jan. 1963, *4*, 44–52.
23. T. Mayer, "The Inflexibility of Monetary Policy," *Rev. Econ. Statist.*, Nov. 1958, *40*, 358–74.
24. J. F. Muth, "Optimal Properties of Exponentially Weighted Forecasts," *J. Amer. Statist. Ass.*, June 1960, *58*, 299–306.
25. M. Nerlove, and K. F. Wallis, "Use of Durbin-Watson Statistic in Inappropriate Situations," *Econometrica*, Jan. 1966, *34*, 235–38.
26. R. M. Solow, "On a Family of Lag Distributions," *Econometrica*, Apr. 1960, *28*, 393–406.
27. J. E. Tanner, "The Timing of the Effects of Monetary Policy," unpublished doctoral thesis, Brown Univ. 1968.
28. J. Tobin, "Liquidity Preference as Behavior Towards Risks," *Rev. Econ. Stud.*, Feb. 1958, *25*, 65–86.
29. D. P. Tucker, "Dynamic Income Adjustment to Money Supply Changes," *Amer. Econ. Rev.*, June 1966, *56*, 433–49.
30. R. N. Waud, "Misspecification in the 'Partial Adjustment' and 'Adaptive Expectations' Models," *Int. Econ. Rev.*, June 1968, *9*, 204–17.

An Empirical Investigation of the Wage-Lag Hypothesis

By THOMAS F. CARGILL*

This paper reports the results of a statistical examination of the wage-lag hypothesis. Despite some questioning, the hypothesis that wages lag behind price changes by a significant amount has been given wide support in economic analysis.[1] This point is clearly illustrated by Clark Warburton in a footnote to a discussion of why wages should lag behind changes in the price level: ". . . Specific references to this aspect of price theory prior to World War I are hardly necessary, since it was so universally accepted" [29, p. 208].

Several important empirical studies have been offered in support of the hypothesis. Most of these have employed graphical analysis of historical time series of prices and money wages, the objective being to determine which series reached a turning point first. A smaller number of studies used regression analysis but have generally been confined to shorter periods of time. Among the most important studies are the following: E. J. Hamilton investigated the relationship between prices, money wages, and industrial development in Spain [15] [16] [17] from 1351 to 1800, in London [13] from 1751 to 1800, and in England [14] from 1500 to 1702. Alvin Hansen [18], W. C. Mitchell [22] [23], Eugene Lerner [20], and R. S. Tucker [26] [27] have examined the relationship between prices and money wages for various periods in the United States. C. Bresciani-Turroni [7] studied the inflationary period in Germany after World War I. All of these investigators used graphical analysis. R. J. Bhatia [4] examined the relationship between wages and profits for the United States during the period from 1935 to 1959 using regression analysis and interpreted his results to imply the existence of a wage lag.

I. *Approach to Determine the Empirical Validity of the Wage-Lag Hypothesis*

The objective of this study is to test the empirical validity of the assumed time relationship between consumer prices and money wages over long periods of time in the United States and England in an attempt to resolve some of the controversy created by A. Alchian and R. Kessel: " . . . In general, it appears that unwarranted validity has been assigned to the wage-lag

* The author is assistant professor of economics at Purdue University. He especially thanks Thomas Mayer for valuable assistance in the preparation of this paper and Charles Schotta for initially arousing an interest in this area of research. In addition, comments from Leon Borgman, Bruce Glassburner, Gordon Rausser, and T. Y. Shen are appreciated. Financial support is acknowledged from the Ford Foundation and UCD computer center. The paper was written while the author was lecturer in economics at the University of California, Davis and is an extension of work in an unpublished dissertation. The author is responsible for remaining errors.

[1] The most important criticisms of both the hypothesis and the studies used to support it were made by A. A. Alchian and R. Kessel [2] and D. Felix [9]. Alchian and Kessel also provided a statistical test of the wage-lag hypothesis using correlation analysis for the period from 1940 to 1952 in the United States. The results of the test were unfavorable. While these criticisms represent a formidable attack on the concept, it is the contention of this study that they do not strike at the core of the hypothesis. This is the assumed time relationship on which the concept stands or falls. The emphasis of these authors is that a wage lag can be accounted for by real forces, and the result of the statistical test by Alchian and Kessel does not directly test for the existence of a stable time relationship between prices and wages.

hypothesis, given the character of the evidence that has been used to support it. A rereading of this evidence suggests that the wage-lag hypothesis ought to be regarded as essentially untested" [2, p. 64]. Alchian and Kessel reached this conclusion after an extensive examination of the most important empirical studies.

The co-movements between prices and wages are analyzed by the application of spectral methods which should provide some meaningful evidence on the empirical value of the hypothesis. Simple regression and visual graphic comparison between prices and wages are not as well suited for this purpose. In the case of regression, data limitations prevent the testing of models with several independent variables, and only a limited number of lag structures can be specified. The more frequently used method of graphical analysis lacks statistical reliability when the only procedure involves the graphing of price and wage series without adjustment for trend and only looking at turning points.

The theory of spectral methods and the estimation procedures will not be reviewed here since they are amply described in other places, [3] [6] [10] [12] [19]; however, there are some basic concepts needing emphasis to interpret the results of this paper. The objective of the spectral approach is to decompose a time-series into frequency components and determine the contribution of each component to the total variance of a series. This is simply a rigorous extension of the traditional decomposition of a time-series into trend, cycles, seasonals, and erratic movement, or long run and short run. To investigate the relationship between two time-series, such as prices and wages, the cross-spectral statistics of *coherence*, *phase*, and *tau* are necessary. *Coherence* indicates the linear association between the price and wage series at the same frequency component and is interpreted in the same manner as R^2 the coefficient of determination. *Phase* is the angular measure of the shift on the time axis of one series relative to another series which maximizes *coherence* at that frequency. This measure records the lead or lag of the price series over the wage series. As a matter of interpretation, when *phase* is positive, the price series leads the wage series at this frequency; otherwise, the wage series leads. Through the use of the *phase* statistic, it is possible to derive the calendar time shift of the series by computing *tau*. *Tau* is equal to the *phase* shift divided by $2\pi f_0$, where f_0 is the frequency in terms of the real time units of the data.[2]

II. *The Data*

The wage-lag hypothesis has been accepted almost as an article of faith by economists and, as such, should hold true for more than several cases. With this in mind, a total of five sets of consumer prices and money wage series are analyzed.[3] There are three sets of yearly time-series for the United States covering the periods 1791–1932, 1820–1965, and 1860–1965. While there is overlapping, there are differences in each of the three sets of data. For example, the money wage series for the period 1791–1932 is based on both daily and weekly wage rates comprising factory earnings, farm wages, building wages, teachers' salaries, etc. The wage series for 1820–1965 is almost entirely based on weekly earnings, whereas the series for the period 1860–1965 is for hourly rates of pay exclusively. These two wage series primarily represent returns to workers in manufacturing pursuits. There are also differences in the price series.

The two sets of data for England cover a

[2] For a recent application of spectral methods employing the three cross-spectral statistics as well as the auto-spectrum, see the study by Thomas J. Sargent [24].

[3] See Appendix.

longer period of time than those for the United States. The first set of prices and wages includes a fairly broad category of workers living in London for the period 1729–1935. Economic historians have always considered London as the center of economic activity in England. The second represents the daily wages received and prices paid by members of the building trades in southern England covering the period 1806–1954. The fact that only building trades are included does not reduce the importance of the empirical results. First, most wage series for England covering long periods of time are usually forced by lack of data to depend heavily on wages paid to builders. Second, for many centuries building workmen constituted a major portion of the labor market.

It is well recognized that each of these historical time-series is subject to the usual criticisms; however, given the quantity and quality of data, they represent the only available material for an empirical investigation of the wage-lag hypothesis over long periods of time. There is, though, one problem that can be partially taken into account. In each case, these sets cover a lengthy period of time during which both the United States and England experienced important structural changes. To make some allowance for this, each set of data is divided into two subperiods, thus resulting in a total of 15 cases to test for the existence of a stable lag of wages behind consumer prices.

III. *Empirical Results*

The spectral method requires the time series to be stationary or essentially trend free. To eliminate the influence of trend characteristic of historical time-series, the spectral results are based on residuals from a log-linear regression.[4] Each set of data and its subperiods is then decomposed into frequency components using the Blackman-Tukey covariance-cosine transformation procedures with the Parzen lag window. The number of lags is 20 for each of the five sets of data and 15 for the subperiods. These allow for an adequate opportunity to determine the timing relationship and at the same time maintain a satisfactory level of degrees of freedom. Several other maximal lags were used with no change in the empirical results.

Tables 1–5 in the Appendix record the cross-spectral statistics of *coherence*, *phase*, and *tau*. These three statistics provide information to estimate the degree of association and the timing relationships between consumer prices and money wages. In the following discussion of these statistics, a convenient distinction is often made between long run (frequencies with periods greater than four years) and short run (frequencies with periods less than four years). While there is a certain amount of arbitrariness in this division, it will serve as a useful framework for discussion. Moreover it points up one of the chief advantages of the spectral approach, namely the examination of a relationship between time-series throughout a wide range of frequencies.

The *coherence* values indicate some interesting results. For the United States, there is a fairly moderate association between prices and wages in the longer run frequencies which is reduced in the shorter run frequencies. The *coherence* values for the data in England also illustrate this pattern; however, the wage and price series for two periods in London and one period in southern England have very low *coherence* values throughout the estimated frequency bands.

The *coherence* values not only indicate

[4] To insure that the results are not appreciably affected by the trend removal method, the data were also analyzed as percentage rates of change and first differences. There was no change in the conclusions of this paper.

the degree of linear association between prices and wages at each frequency component, but are also an inverse index of the variance of the *phase* angle estimate. *Phase* estimates associated with low *coherence* values have little meaning; therefore, only those estimates of *phase* are considered where the hypothesis that the true *coherence* is zero can be rejected at the 95 percent level. At these significant *coherence* values, only the *phase* estimates are considered where the hypothesis that the true *phase* is zero can be rejected at the 95 percent level; otherwise, it is assumed that there is no significant difference in the timing of prices and wages.[5]

For the United States, the low frequency components of the long run have almost all significant *coherence* values, but the majority of the *phase* estimates are not significantly different from zero. The first set of data in Table 1 indicates a wage lag for most of the long-run frequency components during the period 1791–1932 and one subperiod 1860–1932, while the other subperiod shows no significant *phase* shift in these frequencies. In Table 2, only the period 1820–1900 has most of the long-run components showing a wage lag whereas the two periods 1820–1965 and 1900–1965 only have one *phase* shift, each showing a wage lag and wage lead, respectively. The remaining *phase* estimates in the long run are not significantly different from zero. For the results in Table 3, the period 1860–1900 has over half the long components showing a wage lag. The period 1900–1965 has three *phase* shifts showing a wage lead and the period 1860–1965 only two *phase* shifts showing a wage lag. Thus, for the United States, in the long-run frequency components with significant *coherence*, no one particular wage-price timing relationship manifests itself in all of the cases taken together. While several cases indicate the presence of a definite lag of wages there is also evidence of wage leads in one case and no significant *phase* shift between prices and wages in other cases. In the short-run frequencies where *coherence* is significant, there is almost no significant *phase* shift.

Turning our attention to the evidence for England, the 1849–1935 period in London has several significant *coherence* values and *phase* shifts in the long-run frequencies indicating the presence of a wage lag. The other two periods in London together have only one significant *phase* shift showing a wage lag. The long-run frequencies for southern England indicate a wage lag. The short-run frequency components with significant *coherence* values for both London and southern England generally indicate no significant *phase* shift.

If the wage-lag concept is to be given as much confidence as it appears to have been given in the literature, a significant lag in money wages behind prices should have been manifested in the majority of estimated frequencies for each case examined.[6]

[5] The underlying theory of tests of significance, especially for the estimates of *phase*, is still in a state of development and there is not unanimous agreement on the relative merits of the available approaches, for example see C. W. J. Granger and M. Hatanaka [12, p. 80]. The tests used here for the nonzero *coherence* [19, p. 433] and *phase* angle [19, p. 381] depend on the degrees of freedom which varies with the type of lag window used in the estimation procedures. The Parzen lag window allows for the largest number of degrees of freedom for a given record length (n) and number of lags (m). The formula for degrees of freedom is 3.71 n/m. Confidence intervals for the most commonly used lag windows can be found in [5].

[6] The results of this paper have been discussed in terms of significant *coherence* and *phase* estimates; however, there is particular difficulty in the construction of significance tests for *phase* angle and the tests should be regarded as only approximations. Thus it is instructive to examine briefly the results by eliminating the nonzero test for *phase* to determine the effect on the conclusions of this paper. In the long-run frequency components for the United States, the periods 1791–1932 (Table 1), 1791–1860 (Table 1), 1860–1932 (Table 1), 1820–1965 (Table 2), 1820–1900 (Table 2), and 1860–1900 (Table 3) indicate a wage lag; however, the two cases covering the period 1900–1965 (Tables 2 and 3) show a wage lead and the period 1860–1965 (Table 3) shows no definite time relationship. Thus,

In the United States, no one particular timing relationship appears in the long-run frequencies, while in England a wage lag does appear in these components. Even though there appears to be a tendency for wage lags in the low frequencies (long periods), the short-run frequencies tend to exhibit no significant timing difference.

Not only does the wage lag fail to be the sole time relationship between prices and wages, but the existence or nonexistence of a timing relationship for different length cycles is an interesting feature deserving attention. No simple explanation appears readily at hand; however, aided with the intuitive interpretation of the spectral density function [11], a tentative explanation can be suggested. The fact that a timing relationship frequently appears in the long-run frequencies, while prices and wages appear to be coincidental in the short-run frequencies may indicate different factors influencing the price-wage interaction depending on whether one is considering the long or short run. In the long run, the economy is subject to important structural changes which are absent or have little impact in the short run. In fact, the main point of the Kuznets cycle is to emphasize the difference in the underlying process of changes in economic activity between the short business cycle and fluctuations longer than, say, ten years.[7] Over the long run, we can expect significant shifts in the supply of labor to affect the wage level and hence the price-wage interaction. Specifically, shifts in the labor supply can result from changes in population, immigration between countries, and migration from rural to urban areas within the country. As an illustration, assume that an exogenous increase in the money supply results in a sustained growth of aggregate demand and output over a long period of time. This will have a corresponding impact on the demand for labor. Increasing tightness would be reflected in an increase in the rate of change of wages. Price behavior would depend on many factors, but especially the responsiveness of output to demand. A possible result would include prices and wages moving more or less together. However, the price-wage interaction would be modified by changes in the supply of labor via population, immigration and migration changes, whether they are endogenous or exogenous. These forces would make themselves felt for long periods of time and if they implied a timing relationship between prices and wages it should be important and observable. In the short run, the influence of these alterations in the supply of labor is less important and other variables, characteristic of shorter term changes, would underlie price and wage behavior. These include changes in profits, changes in demand associated with inventory fluctuations, changes in raw material prices and agricultural prices. Since these are short term variables, whatever timing relationship resulted between prices and wages would tend to be small in magnitude. The spectral results appear to be consistent with such an interpretation. The lower frequency components of the spectral density function are identified with the longer term movements of prices and wages and measure the influence of events affecting the economy over long periods of time. The higher frequency com-

while the evidence for the wage lag is somewhat stronger, there is still not a complete dominance of this particular timing relationship between prices and wages in the long run. For the short-run frequency components with significant *coherence*, we can reasonably regard prices and wages as coincident given the very small estimates of *phase* in terms of years regardless of the sign of *tau*. In England for the long run, the cases of significant *coherences* indicate a wage lag, but in the short run, the estimates of *tau* are generally small enough to conclude that no meaningful time difference exists between prices and wages. Therefore, even if we eliminate the tests for nonzero *phase*, the wage-lag hypothesis does not find the empirical foundation necessary for the role the hypothesis has played in economic analysis.

[7] See the recent contribution by Moses Abramovitz [1] for a discussion of the factors underlying long term fluctuations in economic activity.

ponents are identified with the shorter run movements in prices and wages thus measuring the influence of events having their impact only for a short period of time. Whatever time relationship resulted be-between prices and wages due to short term factors would necessarily be small and when estimated with the type of data considered in this study may be insignificant.

IV. *Conclusion*

There should be a direct relationship between the confidence given a particular concept or hypothesis in economic analysis and the empirical support offered if theory is to have something useful to say about the real world. There is no need to enumerate the various roles of the wage-lag hypothesis in economic analysis since they are so well known. However, much of the evidence lacks statistical reliability and has been shown by Alchian and Kessel to be unsuitable as a solid foundation for theorizing about the lag of wages behind prices. In this study, similar data as used by previous investigators of the wage-lag hypothesis were subjected to analysis by a modern approach to the study of time series. The results for both the United States and England indicate that there is no agreement with the wage-lag hypothesis in the frequency range identified as the short run since wages and prices appear to be concident here. At the short-run frequency components where there is a meaningful degree of association between prices and wages almost all of the time, differences are not significant. In the long-run frequency range the evidence is somewhat more favorable. The results for the United States show that at the long-run frequency components where *coherence* is significant only about half of the components indicate a significant time difference between prices and wages. At these components, the timing relationship is predominantly a wage lag. However, the remaining time differences are not significant even though there is a meaningful degree of association between prices and wages. At these components, the value of *tau* is generally small. In the case of England where *coherence* is significant, about half of the *phase* shifts show a wage lag and the remainder indicate no significant time difference.

These results argue against the uncritical acceptance of the wage-lag hypothesis in the sense that the wage lag is the sole timing relationship between prices and wages over time. It is this particular interpretation of the hypothesis which has permeated much of the literature. However, the results appear consistent with a less restrictive version of the wage-lag hypothesis. When there is a time difference between prices and wages, it tends to result in a wage lag; however, there are many instances where no meaningful time difference exists, implying that prices and wages are coincident.

(Appendix follows on page 812)

Appendix

Table 1—Cross-Spectral Statistics Between Consumer Prices and Money Wages in the United States 1791–1932

1791–1932				1791–1860				1860–1932			
Period	*Coherence*	*Phase*	*Tau*	Period	*Coherence*	*Phase*	*Tau*	Period	*Coherence*	*Phase*	*Tau*
40.00	.9033[a]	.1378[a]	.8708	30.03	.5936[a]	.0143	.0683	30.03	.9136[a]	.1476[a]	.7046
20.00	.8672[a]	.2341[a]	.7453	14.99	.4659[a]	.0216	.0514	14.99	.9115[a]	.2898[a]	.6920
13.33	.7349[a]	.3055[a]	.6483	10.00	.3970[a]	.1505	.2396	10.00	.8667[a]	.4278[a]	.6809
10.00	.6711[a]	.4050[a]	.6446	7.50	.4708[a]	.3357	.4007	7.50	.8653[a]	.4730[a]	.5647
8.00	.6960[a]	.4245[a]	.5405	6.50	.5307[a]	.3216	.3071	6.50	.8370[a]	.3961[a]	.3782
6.67	.6176[a]	.4317[a]	.4581	5.00	.5785[a]	.1954	.1555	5.00	.7507[a]	.3545[a]	.2821
5.71	.5398[a]	.4029[a]	.3664	4.29	.5244[a]	-.0156	-.0106	4.29	.5841[a]	.3211[a]	.2190
5.00	.5626[a]	.3309[a]	.2633	3.75	.3410[a]	-.2560	-.1528	3.75	.4549[a]	.2298	.1372
4.44	.5486[a]	.0726	.0514	3.33	.1760	-.5211	-.2764	3.33	.6006[a]	.1865	.0989
4.00	.4112[a]	-.2402	-.1529	3.00	.1833	-.6981	-.3333	3.00	.6106[a]	.2348	.1121
3.64	.2056	-.4427	-.2562	2.73	.2300	-.5613	-.2436	2.73	.4604[a]	.3544	.1538
3.33	.0669	-.4887	-.2593	2.50	.1778	-.3108	-.1237	2.50	.3508[a]	.3721	.1481
3.15	.0844	-.2991	-.1465	2.31	.2667	-.1288	-.0473	2.31	.2632	.1688	.0620
2.86	.1462	-.2376	-.1080	2.14	.4070[a]	-.0976	-.0333	2.14	.2141	-.0307	-.0105
2.67	.0951	-.1675	-.0711	2.00	.4330[a]	.0000	.0000	2.00	.1972	.0000	.0000
2.50	.0327	-.0672	-.0267								
2.35	.0702	-.1354	-.0507								
2.22	.2098	-.1109	-.0392								
2.11	.2650	-.0265	-.0089								
2.00	.2624	.0000	.0000								

[a] Significantly different from zero at the 95 percent level.
Phase is measured in radians.
Tau is measured in years and derived from *phase* by dividing *phase* by $2\pi f_0$, where f_0 is the cycles per year.

Table 2—Cross-Spectral Statistics Between Consumer Prices and Money Wages in the United States 1820–1965

1820–1965				1820–1900				1900–1965			
Period	*Coherence*	*Phase*	*Tau*	Period	*Coherence*	*Phase*	*Tau*	*Period*	*Coherence*	*Phase*	*Tau*
40.00	.9320[a]	.0344	.2193	30.03	.8652[a]	.1944[a]	.9280	30.03	.8026[a]	-.1554	-.7419
20.00	.8953[a]	.0654[a]	.2082	14.99	.8159[a]	.3644[a]	.8699	14.99	.8065[a]	-.1868[a]	-.4460
13.33	.7949[a]	.1035	.2196	10.00	.7109[a]	.5043[a]	.8026	10.00	.7242[a]	-.1303	-.2074
10.00	.6885[a]	.1553	.2472	7.50	.6479[a]	.5267[a]	.6287	7.50	.6181[a]	-.0499	-.0595
8.00	.6887[a]	.1283	.1634	6.00	.5637[a]	.4689[a]	.4478	6.00	.6149[a]	-.0034	-.0032
6.67	.7015[a]	.0993	.1054	5.00	.4182[a]	.4031	.3208	5.00	.6773[a]	-.0246	-.0195
5.71	.7147[a]	.0575	.1522	4.29	.3340[a]	.3511	.2395	4.29	.7016[a]	-.0276	-.0188
5.00	.7248[a]	.0071	.0056	3.75	.2458	.2752	.1642	3.75	.6167[a]	-.0515	-.0307
4.44	.7012[a]	-.0269	-.0190	3.33	.2809	.0530	.0281	3.33	.5846[a]	-.0366	-.0194
4.00	.7207[a]	-.0144	-.0092	3.00	.3051	.0481	.0229	3.00	.4944[a]	-.1613	-.0770
3.64	.7815[a]	.0200	.0116	2.73	.2253	.1452	.0630	2.73	.3878[a]	-.0654	-.0284
3.33	.7990[a]	.0481	.0255	2.50	.1348	.1534	.0610	2.50	.3340	.2245	.0893
3.15	.7755[a]	.0578	.0283	2.31	.2280	-.0421	-.0155	2.31	.3498	.3061	.1124
2.86	.7420[a]	.0488	.0222	2.14	.1746	-.1322	-.0451	2.14	.3894[a]	.1560	.0532
2.67	.6802[a]	.0676	.0287	2.00	.1956	.0000	.0000	2.00	.4096[a]	-.0000	-.0000
2.50	.6411[a]	.0763	.0304								
2.35	.6709[a]	.0301	.0113								
2.22	.7060[a]	.0068	.0024								
2.11	.7030[a]	.0126	.0042								
2.00	.6978[a]	.0000	.0000								

For footnotes: see Table 1.

TABLE 3—CROSS-SPECTRAL STATISTICS BETWEEN CONSUMER PRICES AND MONEY WAGES IN THE UNITED STATES 1860–1965

1860–1965				1860–1900				1900–1965			
Period	*Coherence*	*Phase*	*Tau*	Period	*Coherence*	*Phase*	*Tau*	Period	*Coherence*	*Phase*	*Tau*
40.00	.9610[a]	.0148	.0939	30.03	.8157[a]	.3700[a]	1.7667	30.03	.9128[a]	−.0829[a]	−.3956
20.00	.9442[a]	.0082	.0260	14.99	.8026[a]	.4702[a]	1.1225	14.99	.9091[a]	−.1018[a]	−.2430
13.33	.8726[a]	−.0447	−.0948	10.00	.7869[a]	.4578[a]	.7285	10.00	.8928[a]	−.0735[a]	−.1169
10.00	.6990[a]	−.1408	−.2241	7.50	.8366[a]	.3684[a]	.4397	7.50	.8758[a]	−.0324	−.1038
8.00	.6014[a]	−.1274	−.1622	6.00	.8606[a]	.2585[a]	.2468	6.00	.8425[a]	−.0301	−.0287
6.67	.5467[a]	−.0074	−.0079	5.00	.7699[a]	.0629	.0501	5.00	.7850[a]	−.0456	−.0363
5.71	.5900[a]	.0937	.0852	4.29	.5864[a]	−.2356	−.1607	4.29	.7700[a]	−.0244	−.0167
5.00	.5047[a]	.2190	.1743	3.75	.5504[a]	−.2969	−.1772	3.75	.7313[a]	.1038	.0062
4.44	.6121[a]	.3030[a]	.2143	3.33	.6861[a]	−.0617	−.0327	3.33	.6155[a]	−.0450	−.0239
4.00	.7220[a]	.2366[a]	.1506	3.00	.7426[a]	.1523	.0727	3.00	.4904[a]	−.1029	−.0491
3.64	.7429[a]	.1050	.0608	2.73	.7828[a]	.2572	.1116	2.73	.4090[a]	.0675	.0293
3.33	.6974[a]	−.0096	−.0051	2.50	.8283[a]	.2683[a]	.1068	2.50	.4227[a]	.3768	.1449
3.15	.6527[a]	−.0987	−.0483	2.31	.7469[a]	.2105	.0773	2.31	.4915[a]	.2695	.0990
2.86	.6312[a]	−.1011	−.0460	2.14	.5364[a]	.1007	.0343	2.14	.6505[a]	.0155	.0053
2.67	.5810[a]	.0258	.0109	2.00	.4264	.0000	.0000	2.00	.7410[a]	−.0000	−.0000
2.50	.5800[a]	.1465	.0583								
2.35	.6708[a]	.1612	.0604								
2.22	.7531[a]	.1622	.0574								
2.11	.8122[a]	.1228	.0411								
2.00	.8400[a]	.0000	.0000								

For footnotes: see Table 1.

TABLE 4—CROSS-SPECTRAL STATISTICS BETWEEN CONSUMER PRICES AND MONEY WAGES IN ENGLAND (LONDON) 1729–1935

1729–1935				1729–1849				1849–1935			
Period	*Coherence*	*Phase*	*Tau*	Period	*Coherence*	*Phase*	*Tau*	Period	*Coherence*	*Phase*	*Tau*
40.00	.1554	.0466	.2983	30.03	.3178[a]	.3275	1.5644	30.03	.8894[a]	.0680[a]	.3247
20.00	.2652[a]	.1123	.3574	14.99	.3811[a]	.4947[a]	1.1810	14.99	.8715[a]	.1211[a]	.2891
13.33	.4328[a]	.1836	.3897	10.00	.2540[a]	.3492	.5557	10.00	.8857[a]	.1745[a]	.2777
10.00	.4021[a]	.2560	.4074	7.50	.0808	.6006	.7169	7.50	.8572[a]	.2557[a]	.3052
8.00	.2773[a]	.3570	.4546	6.00	.0761	1.2302	1.1748	6.00	.6645[a]	.2423[a]	.2314
6.67	.1197	.5143	.5457	5.00	.0553	1.1103	.8836	5.00	.5681[a]	.1608	.1280
5.71	.0748	.8738	.7947	4.29	.0155	.4835	.3298	4.29	.6770[a]	.1971	.1344
5.00	.0593	.7016	.5583	3.75	.0058	−.1782	−.1064	3.75	.6597[a]	.2998[a]	.1789
4.44	.0702	.2746	.1942	3.33	.0052	.5133	.2723	3.33	.5786[a]	.3252[a]	.1725
4.00	.0474	.4070	.2591	3.00	.0015	−.9009	−.4302	3.00	.5621[a]	.3123[a]	.1491
3.64	.0427	.6764	.3899	2.73	.0114	.4232	.1837	2.73	.5567[a]	.2372	.1030
3.33	.0362	.6905	.3663	2.50	.0496	−1.2738	−.5068	2.50	.5035[a]	.0150	.0060
3.15	.0049	.9584	.4694	2.31	.1981	−.7820	−.2872	2.31	.4233[a]	−.2008	−.0738
2.86	.0051	−.7554	−.3435	2.14	.2139	−.4611	−.1572	2.14	.3240[a]	−.1712	−.0584
2.67	.0003	.5077	.2155	2.00	.1818	.0000	.0000	2.00	.2973[a]	.0000	.0000
2.50	.0381	−.9595	−.3818								
2.35	.0751	−1.0879	−.4074								
2.22	.0641	−1.0851	−.3838								
2.11	.0297	−1.0629	−.3561								
2.00	.0024	.0000	.0000								

For footnotes: see Table 1.

TABLE 5—CROSS-SPECTRAL STATISTICS BETWEEN CONSUMER PRICES AND MONEY WAGES IN ENGLAND (SOUTHERN AREA) 1806–1954

1806–1954				1806–1878				1878–1954			
Period	*Coherence*	*Phase*	*Tau*	Period	*Coherence*	*Phase*	*Tau*	Period	*Coherence*	*Phase*	*Tau*
40.00	.7362[a]	.1009	.6425	30.03	.6805[a]	.1852	.8843	30.03	.6210[a]	.3181[a]	1.5189
20.00	.7218[a]	.1975[a]	.6288	14.99	.5656[a]	.2739[a]	.6538	14.99	.7355[a]	.4564[a]	1.0895
13.33	.7057[a]	.2609[a]	.5535	10.00	.3245[a]	.1658	.2640	10.00	.7083[a]	.5398[a]	.8591
10.00	.5444[a]	.3534[a]	.5624	7.50	.2156	.1089	.1299	7.50	.4981[a]	.7009[a]	.8367
8.00	.3789[a]	.5267[a]	.6706	6.00	.1814	.1328	.1268	6.00	.3889[a]	.6443[a]	.6153
6.67	.2883[a]	.4474	.4747	5.00	.1028	.3826	.3044	5.00	.4444[a]	.3816	.3037
5.71	.3136[a]	.2497	.2271	4.29	.0902	1.0662	.7272	4.29	.4872[a]	.1113	.0759
5.00	.4195[a]	.1601	.1274	3.75	.1160	1.2148	.7244	3.75	.4887[a]	−.1735	−.1035
4.44	.4867[a]	.0891	.0630	3.33	.0627	1.1495	.6098	3.33	.5106[a]	−.2524	−.1339
4.00	.4481[a]	.0510	.0325	3.00	.0142	−.8126	−.3880	3.00	.5904[a]	−.2025	−.0967
3.64	.4234[a]	.0492	.0285	2.73	.0632	.0192	.0083	2.73	.6676[a]	−.1654	−.0718
3.33	.5230[a]	−.0802	−.0425	2.50	.0500	.6049	.2407	2.50	.6733[a]	−.1935	−.0770
3.15	.5205[a]	−.1909	−.0935	2.31	.0996	−1.2092	−.4441	2.31	.6010[a]	−.2248	−.0826
2.86	.6319[a]	−.1862	−.0847	2.14	.1382	−.7588	−.2588	2.14	.5075[a]	−.1824	−.0622
2.67	.6229[a]	−.1819	−.0772	2.00	.0701	.0000	.0000	2.00	.4648[a]	.0000	.0000
2.50	.6160[a]	−.2114	−.0841								
2.35	.5843[a]	−.1583	−.0593								
2.22	.5269[a]	−.0893	−.0316								
2.11	.4937[a]	−.0880	−.0295								
2.00	.4959[a]	.0000	.0000								

For footnotes: see Table 1.

Each of these series is an index based on direct price and wage observations from numerous sources. No relationships between the series and other variables like industrial output were used to derive the various series.

The following are the data sources for the three sets of consumer prices and wages for the United States: (1) 1791–1932: Tucker has estimated the series for consumer prices [26] and wage rates [27] covering the entire period. Tucker's estimate of the price series is composed of two parts. The first is his own estimate for the period 1791–1860 on the assumption that a cost of living series from retail prices corresponds fairly closely with a simple average of wholesale prices and money wages. For the second period, Tucker added a price series estimated by Carl Snyder [25], which includes wholesale prices, retail prices, rents, real estate, and other items. The money wage series is based on data for both daily and weekly rates of pay. Before 1841, Tucker based the wage series on estimates of wages paid to workers living in Massachusetts from 1840–1890, he averaged together the wage series estimated by Hansen [18] and Snyder [25]; and for the remainder of the period used the Snyder series. (2) 1820–1965: The consumer price series for the entire period is obtained from the Bureau of Labor Statistics (*BLS*) [31, p. 199] and represents the prices paid by city workers for all items, including rent, apparel, and upkeep. The money wage series represents primarily weekly wages. For the period 1820–1923, the estimates of money wages provided by Hansen [18] are used. Hansen's series is derived by joining together several wage series estimated by R. P. Falkner, Wesley Mitchell, Paul Douglas and F. Lamberson and others. The money wage series is brought up to 1965 with the *BLS* estimates [31, p. 150] of average weekly earnings of production workers in manufacturing. (3) 1860–1965: Both the consumer price series and money wage series are provided by C. D. Long [21] for the period 1960–1958 and brought up to 1965 with estimates by the *BLS* [30]. The consumer price series for this period is essentially the same as used in (2) except for a few short sub-

periods. The money wage series represents hourly wages based on Long's own estimates as well as estimates made by the *BLS* and Albert Rees.

The following are the sources for the two sets of consumer prices and money wages for England: (1) 1729–1935: Tucker [28] has derived estimates of both the consumer price and money wage series for the entire period in London. The consumer price series is based on a fairly broad range of commodities. In addition, Tucker adds an estimate of the amount of rent paid by workers living in London on the assumption that one-sixth of a worker's weekly earnings went for rent. The wage series represents weekly earnings. For the first several decades, much reliance is placed on workers in the building trades given the lack of data; however, after 1860, it was possible to include far more trades. (2) 1806–1954: The price and wage series for this entire period are those of E. H. Phelps Brown and S. Hopkins [8] who have estimated daily wages and prices for building trades in the southern part of England for the entire period. Unlike the price series for the United States, the Brown-Hopkins price series includes no estimate for rent for the entire period. A straight line approximation for two missing years, 1865 and 1872, was made by the author in the wage series for this period. The Brown-Hopkins series goes back to the thirteenth century; however, the earlier data was not used because of dubious reliability and the fact that there were few changes in the money wage index up to about 1800.

References

1. M. Abramovitz, "The Passing of the Kuznets Cycle," *Economica*, Feb. 1969, *36*, 349–67.
2. A. A. Alchian and R. A. Kessel, "The Meaning and Validity of the Inflation Induced Lag of Wages Behind Prices," *Amer. Econ. Rev.*, Mar. 1960, *50*, 43–66.
3. J. S. Bendat and A. G. Piersol, *Measurement and Analysis of Random Data*, New York 1966.
4. R. H. Bhatia, "Profits and the Rate of Change of Money Earnings in the United States," *Economica*, Aug. 1962, *29*, 255–62.
5. V. Bonomo and C. Schotta, *Statistical Tables for Time Series Analysis*, Technical Paper No. 69-1, Virginia Polytechnic Institute, Jan. 1969.
6. L. E. Borgman, *Statistical Time Series*, Berkeley 1967.
7. C. Bresciani-Turroni, *The Economics of Inflation*, London 1937.
8. E. H. Phelps Brown and S. V. Hopkins, "Seven Centuries of the Prices of Consumables Compared with Builders' Wage-Rates," *Economica*, Nov. 1956, *23*, 296–314.
9. D. Felix, "Profit Inflation and Industrial Growth: The Historical Record and Contemporary Analogies," *Quart. J. Econ.*, Aug. 1956, *70*, 441–63.
10. G. S. Fishman, *Spectral Methods in Econometrics*, RAND Corporation Report R-453-PR, Santa Monica 1968.
11. C. W. J. Granger, "The Typical Spectral Shape of an Economic Variable," *Econometrica*, Jan. 1966, *34*, 150–61.
12. ——— and M. Hatanaka, *Spectral Analysis of Economic Time Series*, Princeton 1964.
13. E. J. Hamilton, "Profit Inflation and the Industrial Revolution, 1751–1800," *Quart. J. Econ.*, Feb. 1942, *56*, 256–73; reprinted in F. C. Lane and J. C. Riemersma, eds., *Enterprise and Secular Change*, Homewood 1953.
14. ———, "American Treasure and the Rise of Capitalism (1500–1700)," *Economica*, Nov. 1929, *9*, 338–57.
15. ———, *Money, Prices and Wages in Valencia, Aragon, and Navarre, 1351–1500*, Cambridge, Mass. 1936.
16. ———, *American Treasure and the Price Revolution in Spain, 1501–1650*, Cambridge, Mass. 1934.
17. ———, *War and Prices in Spain, 1651–1800*, Cambridge, Mass. 1947.
18. A. H. Hansen, "Factors Affecting the Trend of Real Wages," *Amer. Econ. Rev.*, Mar. 1925, *15*, 27–42.
19. G. M. Jenkins and D. G. Watts,

Spectral Analysis and its Applications, San Francisco 1968.
20. E. M. LERNER, "Money, Prices, and Wages in the Confederacy, 1861–65," *J. Polit. Econ.*, Feb. 1955, *63*, 20–40.
21. C. D. LONG, "The Illusion of Wage Rigidity: Long and Short Cycles in Wages and Labor," *Rev. Econ. Statist.*, May 1960, *42*, 140–51.
22. W. C. MITCHELL, *Gold, Prices, and Wages under the Greenback Standard*, Berkeley 1908.
23. ———, *A History of the Greenbacks*, Chicago 1903.
24. T. J. SARGENT, "Interest Rates in the Nineteen-Fifties," *Rev. Econ. Statist.*, May 1968, *50*, 164–72.
25. C. SNYDER, "The Measurement of the General Price Level," *Rev. Econ. Statist.*, Feb. 1928, *10*, 40–52.
26. R. S. TUCKER, "Gold and the General Price Level," *Rev. Econ. Statist.*, Jan-Feb. 1934, *16*, 8–16 and 25–27.
27. ———, "Real Wages Under Laissez-Faire," *Barron's, The National Financial Weekly*, No. 43, Oct. 23, 1933, 6–7.
28. ———, "Real Wages of Artisans in London, 1729–1935," *J. Amer. Statist. Ass.*, Mar. 1936, *31*, 73–84.
29. C. WARBURTON, "The Misplaced Emphasis in Contemporary Business Fluctuation Theory," *J. Bus., Univ. Chicago*, Oct. 1946, *19*, 199–220; reprinted in F. A. Lutz and L. W. Mints, eds., *Readings in Monetary Theory*, Homewood 1951.
30. COUNCIL OF ECONOMIC ADVISORS, *Economic Report of the President*, Jan. 1967.
31. U.S. DEPARTMENT OF LABOR, Bureau of Labor Statistics, *Handbook of Labor Statistics, 1967*, Bull 1555, Washington 1967.

An Economic Interpretation of Optimal Control Theory

By ROBERT DORFMAN*

Capital theory is the economics of time. Its task is to explain if, and why, a lasting instrument of production can be expected to contribute more to the value of output during its lifetime than it costs to produce or acquire. From the explanation, it deduces both normative and descriptive conclusions about the time-path of the accumulation of capital by economic units and entire economies.

Traditionally, capital theory, like all other branches of economics, was studied in the context of stationary equilibrium. For example, the stationary state of the classical economists, and the equilibrium of Böhm-Bawerk's theory of the period of production, both describe the state of affairs in which further capital accumulation is not worthwhile. A mode of analysis that is confined to a distant, ultimate position is poorly suited to the understanding of accumulation and growth,[1] but no other technique seemed available for most of the history of capital theory.

For the past fifty years it has been perceived, more or less vaguely, that capital theory is formally a problem in the calculus of variations.[2] But the calculus of variations is regarded as a rather arcane subject by most economists and, besides, in its conventional formulations appears too rigid to be applied to many economic problems. The application of this conceptual tool to capital theory remained peripheral and sporadic until very recently, and capital theory remained bound by the very confining limitations of the ultimate equilibrium.

All this has changed abruptly in the past decade as a result of a revival, or rather reorientation, of the calculus of variations prompted largely by the requirements of space technology.[3] In its modern version, the calculus of variations is called optimal control theory. It has become, deservedly, the central tool of capital theory and has given the latter a new lease on life. As a result, capital theory has become so profoundly transformed that it has been rechristened growth theory, and has come to grips with numerous important practical and theoretical issues that previously could not even be formulated.

The main thesis of this paper is that optimal control theory is formally identical with capital theory, and that its main insights can be attained by strictly economic reasoning. This thesis will be supported by deriving the principal theorem of optimal control theory, called the maximum principle, by means of economic analysis.

I. *The Basic Equations*

In order to have a concrete vocabulary, consider the decision problem of a firm that wishes to maximize its total profits over some period of time. At any date t, this firm will have inherited a certain stock of capital and other conditions from its

* The author is professor of economics at Harvard University.

[1] A point made most forceably by Joan Robinson in [9] and elsewhere.

[2] Notable examples are Hotelling [6] and Ramsey [8].

[3] The twin sources of the new calculus of variations are R. Bellman [4] and L. S. Pontryagin, et al. [7]. Bellman emphasized from the first the implications of his work for economics.

past behavior. Denote these by $k(t)$. With this stock of capital and other facilities k and at that particular date t, the firm is in a position to take some decisions which might concern rate of output, price of output, product design, or whatnot. Denote the decisions taken at any date by $x(t)$. From the inherited stock of capital at the specified date together with the specified current decisions the firm derives a certain rate of benefits or net profits per unit of time. Denote this by $u(k(t), x(t), t)$.[4] This function u determines the rate at which profits are being earned at time t as a result of having k and taking decisions x.

Now look at the situation as it appears at the initial date $t=0$. The total profits that will be earned from then to some terminal date T is given by:

$$W(k_0, \vec{x}) = \int_0^T u(k, x, t)dt$$

which is simply the sum of the rate at which profit is being earned at every instant discounted to the initial date (if desired) and added up for all instants.[5] In this notation $\vec{x}$ does not denote an ordinary number but the entire time path of the decision variable x from the initial date to T. This notation asserts that if the firm starts out with an initial amount of capital k_0 and then follows the decision policy denoted by $\vec{x}$, it will obtain a total result, W, which is the integral of the results obtained at each instant; these results in turn depending upon the date of the pertinent instant, the capital stock then and the decision applicable to that moment. The firm is at liberty, within limits, to choose the time path of the decision variable $\vec{x}$ but it cannot choose independently the amount of capital at each instant; that is a consequence of the capital at the initial date and the time path chosen for decision variable. This constraint is expressed by saying that the rate of change of the capital stock at any instant is a function of its present standing, the date, and the decisions taken. Symbolically:[6]

$$(1) \qquad \dot{k} = \frac{dk}{dt} = f(k, x, t).$$

Thus the decisions taken at any time have two effects. They influence the rate at which profits are earned at that time and they also influence the rate at which the capital stock is changing and thereby the capital stock that will be available at subsequent instants of time.

These two formulas express the essence of the problem of making decisions in a dynamic context. The problem is to select the time path symbolized by $\vec{x}$ so as to make the total value of the result, W, as great as possible taking into account the effect of the choice of x on both the instantaneous rate of profit and the capital stock to be carried into the future. This is truly a difficult problem, and not only for beginners. The essential difficulty is that an entire time path of some variable has to be chosen. The elementary calculus teaches how to choose the best possible number to assign to a single variable or the best numbers for a few variables by differentiating some function and setting partial derivatives equal to zero. But finding a best possible time path is an entirely different matter and leads into some very advanced mathematics. The strategy of the solution is to reduce the problem which, as it stands, requires us to find an entire time path, to a problem which demands us to determine only a single number (or a few numbers), which is something we know how to do from the ordinary cal-

[4] In the sequel we shall often omit the time-arguments in the interest of simplicity, and thus write simply $u(k, x, t)$.

[5] The argument t allows the introduction of any discounting formula that may be appropriate.

[6] The dot will be used frequently to denote a rate of change with respect to time.

culus. This transformation of the problem can be performed in a number of ways. One way, which dates back to the eighteenth century, leads to the classical calculus of variations. Another way, which will be followed here, leads to the maximum principle of optimal control theory. This method depends very heavily on introducing the proper notation. First, introduce a formula for the value that can be obtained by the firm starting at an arbitrary date t with some amount of capital k and then following an arbitrary decision policy $\vec{x}$ until the terminal date. It is

$$W(k_t, \vec{x}, t) = \int_t^T u[k, x, \tau]d\tau$$

which, of course, is just a generalization of the W formula introduced previously.

Now break W up into two parts. Think of a short time interval of length Δ beginning at time t. Δ is to be thought of as being so short that the firm would not change x in the course of it even if it could. Then we can write

$$W(k, \vec{x}, t) = u(k, x_t, t)\Delta + \int_{t+\Delta}^T u[k(t), x, \tau]d\tau. \tag{2}$$

This formula says that if the amount of capital available at time t is k and if the policy denoted by $\vec{x}$ is followed from then on, then the value contributed to the total sum from date t on consists of two parts. The first part is the contribution of a short interval that begins at date t. It is the rate at which profits are earned during the interval times the length of the interval. It depends on the current capital stock, the date, and the current value of the decision variable, here denoted by x_t. The second part is an integral of precisely the same form as before but beginning at date t+Δ. It should be noticed that the starting capital stock for this last integral is not k(t) but k(t+Δ). This fact, that the capital stock will change during the interval in a manner influenced by x_t, will play a very significant role. We can take advantage of the fact that the same form of integral has returned by writing

$$W(k_t, \vec{x}, t) = u(k, x_t, t)\Delta + W(k_{t+\Delta}, \vec{x}, t + \Delta)$$

where the changes in the subscripts are carefully noted.

Now some more notation. If the firm knew the best choice of $\vec{x}$ from date t on, it could just follow it and thereby obtain a certain value. We denote this value, which results from the optimal choice of $\vec{x}$ by V^*, as follows

$$V^*(k_t, t) = \max W(k_t, \vec{x}, t).$$

Notice that V^* does not involve $\vec{x}$ as an argument. This is because $\vec{x}$ has been maximized out. The maximum value that can be obtained beginning at date t with capital k does not depend on $\vec{x}$ but is the value that can be obtained in those conditions from the best possible choice of $\vec{x}$. Now suppose that the policy designated by x_t is followed in the short time interval from t to t+Δ and that thereafter the best possible policy is followed. By formula (2) the consequence of this peculiar policy can be written as

$$V(k_t, x_t, t) = u(k_t, x_t, t)\Delta + V^*(k_{t+\Delta}, t + \Delta).$$

In words, the results of following such a policy are the benefits that accrue during the initial period using the decision x_t plus the maximum possible profits that can be realized starting from date t+Δ with capital k(t+Δ) which results from the decision taken in the initial period.

Now we have arrived at the ordinary calculus problem of finding the best possible value for x_t. If the firm adopts this value, then V of the last formula will be equal to V^*. The calculus teaches us that one frequently effective way to discover a value of a variable that maximizes a given function is to differentiate the function

with respect to the variable and equate the partial derivative to zero. This is the method that we shall use. But first we should be warned that this method is not sure-fire. It is quite possible for the partial derivatives to vanish when the function is not maximized (for example, they may vanish when it is minimized), and cases are not rare in which the partial derivatives differ from zero at the maximum. We shall return to these intricacies later. For the present we assume that the partial derivative vanishes at the maximum, differentiate $V(k_t, x_t, t)$ with respect to x_t, and obtain

$$\Delta \frac{\partial}{\partial x_t} u(k, x_t, t) + \frac{\partial}{\partial x_t} V^*(k(t+\Delta), t+\Delta) = 0. \tag{3}$$

The trouble with that formula, aside from the fact that the function V^* is still unknown, is that we are told to differentiate V^* with respect to x_t, whereas it does not involve x_t explicitly. To get around this, notice

$$\frac{\partial V^*}{\partial x_t} = \frac{\partial V^*}{\partial k(t+\Delta)} \frac{\partial k(t+\Delta)}{\partial x_t}.$$

Both of these expressions merit some analysis and we shall start with the second. Since we are dealing with short time periods we can use the approximation

$$k(t+\Delta) = k(t) + \dot{k}\Delta.$$

That is, the amount of capital at $t+\Delta$ is equal to the amount of capital at t plus the rate of change of capital during the interval times the length of the interval. Remembering formula (1), $\dot{k}$ depends on x_t:

$$\dot{k} = f(k, x_t, t).$$

Thus we can write

$$\frac{\partial k(t+\Delta)}{\partial x_t} = \Delta \frac{\partial f}{\partial x_t}.$$

Turn, now, to the first factor, $\partial V^*/\partial k$. This derivative is the rate at which the maximum possible profit flow from time $t+\Delta$ on changes with respect to the amount of capital available at $t+\Delta$. It is, therefore, the marginal value of capital at time $t+\Delta$, or the amount by which a unit increment in capital occurring at that time would increase the maximum possible value of W. We denote the marginal value of capital at time t by $\lambda(t)$, defined by

$$\lambda(t) = \frac{\partial}{\partial k} V^*(k, t).$$

Inserting these results in formula (3), we obtain

$$\Delta \frac{\partial u}{\partial x_t} + \lambda(t+\Delta)\Delta \frac{\partial f}{\partial x_t} = 0 \tag{4}$$

and furthermore, the constant Δ can be cancelled out. We have one more simplification to make before arriving at our first important conclusion. The marginal value of capital changes gradually over time and so, to a sufficiently good approximation,

$$\lambda(t+\Delta) = \lambda(t) + \dot{\lambda}(t)\Delta.$$

That is, the marginal value of capital at $t+\Delta$ is the marginal value at t plus the rate at which it is changing during the interval multiplied by the length of the interval. Insert this expression in equation (4), after cancelling the common factor Δ in the equation as written, to obtain

$$\frac{\partial u}{\partial x_t} + \lambda(t)\frac{\partial f}{\partial x_t} + \dot{\lambda}(t)\Delta \frac{\partial f}{\partial x_t} = 0.$$

Now allow Δ to approach zero. The third term becomes negligibly small in comparison with the other two. Neglecting it, there results:

$$\frac{\partial u}{\partial x_t} + \lambda \frac{\partial f}{\partial x_t} = 0. \tag{5}$$

This is our first major result and con-

stitutes about half of the maximum principle. It makes perfectly good sense to an economist. It says that along the optimal path of the decision variable at any time the marginal short-run effect of a change in decision must just counter-balance the effect of that decision on the total value of the capital stock an instant later. We see that because the second term in the equation is the marginal effect of the current decision on the rate of growth of capital with capital valued at its marginal worth, λ. The firm should choose x at every moment so that the marginal immediate gain just equals the marginal long-run cost, which is measured by the value of capital multiplied by the effect of the decision on the accumulation of capital.

Now suppose that x_t is determined so as to satisfy equation (5). On the assumption that this procedure discovers the optimal value of x_t, $V(k_t, x_t, \mathrm{t})$ will then be equal to its maximum possible value or $V^*(k, \mathrm{t})$. Thus

$$V^*(k, \mathrm{t}) = u(k, x_t, \mathrm{t})\Delta + V^*(k(\mathrm{t}+\Delta), \mathrm{t}+\Delta).$$

Now differentiate this expression with respect to k. The derivative of the left-hand side is by definition $\lambda(\mathrm{t})$. The differentiation of the right-hand side is very similar to the work that we have already done and goes as follows:

$$\begin{aligned}\lambda(\mathrm{t}) &= \Delta \frac{\partial u}{\partial k} + \frac{\partial}{\partial k} V^*(k(\mathrm{t}+\Delta), \mathrm{t}+\Delta)\\ &= \Delta \frac{\partial u}{\partial k} + \frac{\partial k(\mathrm{t}+\Delta)}{\partial k} \lambda(\mathrm{t}+\Delta)\\ &= \Delta \frac{\partial u}{\partial k} + \left(1 + \Delta \frac{\partial f}{\partial k}\right)(\lambda + \dot{\lambda}\Delta)\\ &= \Delta \frac{\partial u}{\partial k} + \lambda + \Delta\lambda \frac{\partial f}{\partial k} + \Delta \dot{\lambda} + \dot{\lambda}\frac{\partial f}{\partial k}\Delta^2.\end{aligned}$$

We can ignore the term in Δ^2 and make the obvious cancellations to obtain

$$(6) \qquad -\dot{\lambda} = \frac{\partial u}{\partial k} + \lambda \frac{\partial f}{\partial k}.$$

This is the second major formula of the maximum principle and possesses an illuminating economic interpretation.

To a mathematician, $\dot{\lambda}$ is the rate at which the value of a unit of capital is changing. To an economist, it is the rate at which the capital is appreciating. $-\dot{\lambda}(\mathrm{t})$ is therefore the rate at which a unit of capital depreciates at time t. Accordingly the formula asserts that when the optimal time path of capital accumulation is followed, the decrease in value of a unit of capital in a short interval of time is the sum of its contribution to the profits realized during the interval and its contribution to enhancing the value of the capital stock at the end of the interval. In other words, a unit of capital loses value or depreciates as time passes at the rate at which its potential contribution to profits becomes its past contribution.

This finding is reminiscent of the figure of speech employed by the nineteenth century capital theorists. They said that a capital good embodied a certain amount of value which it imparted gradually to the commodities that were made with its assistance. That is just what is going on here. Each unit of the capital good is gradually decreasing in value at precisely the same rate at which it is giving rise to valuable outputs, either currently saleable or stored for the future in accumulated capital. We can also interpret $-\dot{\lambda}$ as the loss that would be incurred if the acquisition of a unit of capital were postponed for a short time.

II. *The Maximum Principle*

In effect we have been led to construct the auxiliary or Hamiltonian function

$$H = u(k, x, \mathrm{t}) + \lambda(\mathrm{t})f(k, x, \mathrm{t}),$$

to compute its partial derivative with re-

spect to x, and to set that partial derivative equal to zero. This construction has substantial economic significance. If we imagine H to be multiplied by Δ, we can see that it is the sum of the total profits earned in the interval Δ plus the accrual of capital during the interval valued at its marginal value. $H\Delta$ is thus the total contribution of the activities that go on during the interval Δ, including both its direct contribution to the integral W, and the value of the capital accumulated during the interval. Naturally, then, the decision variable x during the current interval should be chosen so as to make H as great as possible. It is for this reason that the procedure we are describing is called the maximum principle. A simple and frequently effective way to do this is to choose a value of the control variable for which the partial derivative vanishes, as we have done.

We have also, in effect, computed the partial derivative of H with respect to k and equated that partial derivative to $-\dot{\lambda}$. The common sense of this operation can be seen best from a modified Hamiltonian,

$$
\begin{aligned}
H^* &= u(k, x, t) + \frac{d}{dt}\lambda k \\
&= u(k, x, t) + \lambda \dot{k} + \dot{\lambda} k.
\end{aligned}
$$

$H^*\Delta$ is the sum of the profits realized during an interval of length Δ and the increase in the value of the capital stock during the interval, or in a sense, the value of the total contribution of activities during the interval to current and future profits.[7] If we maximize H^* formally with respect to x and k we obtain:

$$
\frac{\partial u}{\partial x} + \lambda \frac{\partial f}{\partial x} = 0,
$$

$$
\frac{\partial u}{\partial k} + \lambda \frac{\partial f}{\partial k} + \dot{\lambda} = 0,
$$

[7] H^* differs from H by including capital gains.

which are equations (5) and (6).

Of course, the firm cannot maximize H^* with respect to k since k is not a variable subject to choice. But we now see that equations (5) and (6) advise the firm to choose the time-paths of x and λ so that the resultant values of k are the ones it would choose, if it could do so, to make the sum of profits and increment in capital value as great as possible in every short time interval.

As a technical note, in differentiating H, the marginal value λ is not regarded as a function of x and k, but as a separate time path which is to be determined optimally.

Now we have before us the basic ideas of the maximum principle. There is naturally much more to the method than these two formulas. A good deal of mathematical elaboration is required before the two formulas can be implemented, and we shall indicate later some of the complications that can arise. But there is one additional feature that has to be mentioned before we have finished dealing with fundamentals. This concerns the boundary conditions; for example, the amount of capital available at the beginning of the planning period and the amount required to be on hand on the terminal date.

To see how these boundary data affect the solution to the problem, consider how the three basic formulas operate. They are:

$$
\text{(I)} \qquad \dot{k} = f(k, x, t)
$$

$$
\text{(II)} \qquad \frac{\partial u}{\partial x} + \lambda \frac{\partial f}{\partial x} = 0
$$

$$
\text{(III)} \qquad \frac{\partial u}{\partial k} + \lambda \frac{\partial f}{\partial k} = -\dot{\lambda}.
$$

The first of these is part of the data of the problem. It specifies how capital grows at any instant as a result of its current standing and the choices made. The other two formulas are the main results of the maximization principle. Formula II says that

the choice variable at every instant should be selected so that the marginal immediate gains are in balance with the value of the marginal contribution to the accumulation of capital. Formula III says that capital depreciates at the same rate that it contributes to useful output.

The three formulas are conveniently written and remembered in terms of the Hamiltonian. In this form they are:

$$\text{(I}'\text{)} \qquad \frac{\partial H}{\partial \lambda} = \dot{k}$$

$$\text{(II}'\text{)} \qquad \frac{\partial H}{\partial x} = 0$$

$$\text{(III}'\text{)} \qquad \frac{\partial H}{\partial k} = -\dot{\lambda}.$$

Notice the reciprocal roles played by k and λ in these equations. The partial derivative of H with respect to either is simply related to the time-derivative of the other.

These three formulas jointly determine completely the time paths of the choice variable, the capital stock, and the value of capital. We shall start at time zero with a certain capital stock and a certain initial value for capital. Now look at formula II written out a bit more explicitly:

$$\text{(II)} \qquad \frac{\partial}{\partial x} u(k, x, t) + \lambda(t) \frac{\partial}{\partial x} f(k, x, t) = 0.$$

With k and λ known, this formula determines the value of x, the choice variable.[8] Putting this value in formula I we obtain $\dot{k}$, the rate at which the capital stock is changing. Putting it in formula III we similarly obtain $\dot{\lambda}$ the rate at which the value of a unit of capital is changing. Thus we know the capital stock and the value of a unit of capital a short time later. Using these new values, we can repeat our substitutions in the three formulas and so find, in order, a new value of the choice variable, a new rate for the change in the capital stock and a new rate for the change in the value of capital. Repeating this cycle over and over again, we can trace through the evolution of all the variables from time zero to time T.

In short, these three formulas working together determine the optimal paths of all the variables starting out from any given initial position. In another sense, then, the problem of the choice of an optimal path has been reduced to a much simpler problem, the problem of choosing an optimal initial value for the value of a unit of capital. This is not by any means an easy problem, but it is obviously a great deal easier than finding an entire optimal path without the aid of these formulas.

III. *The Boundary Conditions*

We can now mention the role of boundary conditions. They are of two sorts. Initial conditions describe the state of the firm or economy at the initial date, t=0. In particular, they set forth the initial stock of capital. Terminal conditions prescribe the values of some, or all, of the variables at the terminal date, t=T. For example, the problem may require that the firm have at least some specified stock of capital, say $\bar{K}$ on hand at the terminal date, which can be imposed by including $k(T) \geq \bar{K}$ in the conditions of the problem. Or, again, if the problem is strictly one of maximizing profits during a finite interval, 0 to T, it is clear that capital on hand at date T cannot contribute to that objective; it exists too late to be of service before date T. Such a problem gives rise to the terminal condition $\lambda(T) = 0$.

Now we have seen that the three equations (I), (II), (III) jointly determine the entire evolution of x, k, and λ once the starting values have been prescribed. In particular they determine the terminal

[8] Some mathematical complications arise here. We assume that with k, λ, and t given, formula (II) is satisfied by a unique value of x.

values. We have only[9] to determine a set of starting values that leads to acceptable terminal values to find an entire time path that satifies the necessary conditions for being optimal. In our example, since the initial capital stock is given, the critical initial value to be determined is $\lambda(0)$, the marginal value of capital at the initial date. The three basic formulas, abstract though they may appear, in fact constitute a constructive solution to the problem of choosing an optimal time path. They are a solution, in principle, of the problem of optimal capital accumulation.

We have now found that the old-fashioned technique of equating margins, used with a little ingenuity, leads to the maximum principle, which is the fundamental theorem of optimal control theory.

IV. *An Example*

About the simplest known example of the application of these principles to an economic problem is the derivation of the socially optimal path of capital accumulation for a one-sector economy with an exponentially growing population and production under constant returns to scale.[10]

Let us set forth some notation and data. $N(\mathrm{t})$ is population at date t. Since population grows exponentially, at rate n, say,

$$N(\mathrm{t}) = N(0)e^{n\mathrm{t}}.$$

It will save clutter if we assume $N(0)=1$ (measured in hundreds of millions of people). Denote per capita consumption by c and the utility enjoyed by a person consuming at rate c by $u(c)$. The total utility enjoyed by all the persons alive at time t with per capita consumption at rate c is

$$e^{n\mathrm{t}}u(c).$$

Let ρ be the social rate of time preference. Then the importance at time 0 of the consumption achieved at time t is

$$e^{-\rho \mathrm{t}}e^{n\mathrm{t}}u(c) = e^{(n-\rho)\mathrm{t}}u(c). \tag{7}$$

A defensible social objective for a society with time horizon T (conceivably infinite) is to maximize

$$W = \int_0^{\mathrm{T}} e^{(n-\rho)\mathrm{t}}u(c)\, d\mathrm{t}, \tag{8}$$

or the sum of the utilities enjoyed between 0 and T.[11]

Consumption is limited by output and output by capital stock. Let $K(\mathrm{t})$ denote the capital stock at date t and let $k(\mathrm{t}) = K(\mathrm{t})/N(\mathrm{t})$ denote capital per capita. By virtue of constant returns to scale, we can write the production function of the economy as

$$Y(\mathrm{t}) = N(\mathrm{t})f(k(\mathrm{t}))$$

or, omitting the confusing time-arguments,

$$Y = Nf(k) = e^{n\mathrm{t}}f(k).$$

Gross investment equals output minus consumption, or Y-Nc. Net investment equals gross investment minus physical depreciation. Suppose that physical capital deteriorates at the rate δ per unit per annum so that the total rate of decay of the physical stock, when it is K, is δK. Then net capital accumulation is

$$\begin{aligned} \dot{K} &= Y - Nc - \delta K = N(f(k) - c) - \delta K \\ &= N(f(k) - c) - \delta Nk \\ &= N(f(k) - c - \delta k). \end{aligned}$$

Finally, eliminate $\dot{K}$ by noticing:

$$\begin{aligned} \dot{k} &= \frac{d}{d\mathrm{t}}\frac{K}{N} = \frac{K}{N}\left(\frac{\dot{K}}{K} - \frac{\dot{N}}{N}\right) \\ &= k\left(\frac{\dot{K}}{Nk} - n\right) \qquad (9) \\ &= f(k) - c - \delta k - nk \\ &= f(k) - c - (n+\delta)k. \end{aligned}$$

[9] Only! Reputations have been made by solving this problem in important instances.

[10] A more extended discussion of a very similar model can be found in Arrow [1].

[11] It is best to assume $\rho > n$ or else the integral will be infinite for $\mathrm{T} = \infty$.

Equations (8) and (9) constitute our simple example. Equation (9) is an example of equation (I). To derive equation (II), differentiate equations (7) and (9) with respect to the choice variable, c:

$$\frac{\partial}{\partial c} e^{(n-\rho)t}u(c) = e^{(n-\rho)t}u'(c),$$

$$\frac{\partial}{\partial c} [f(k) - c - (n+\delta)k] = -1.$$

Hence equation (II) is:

$$e^{(n-\rho)t}u'(c) - \lambda = 0, \tag{10}$$

or the value of a unit of capital at time t is the marginal utility of consumption at that time, adjusted for population growth and the social rate of time preference.

Equation (III) is obtained similarly by differentiating equations (7) and (9) with respect to k. There results:

$$-\dot{\lambda} = 0 + \lambda[f'(k) - (n+\delta)],$$

or

$$f'(k) = n + \delta - \frac{\dot{\lambda}}{\lambda}\cdot \tag{11}$$

Equation (10) can be used to eliminate the unfamiliar λ. Differentiating it with respect to time:

$$\frac{\dot{\lambda}}{\lambda} = n - \rho + \frac{u''(c)}{u'(c)}\frac{dc}{dt}\cdot$$

Thus equation (11) becomes

$$f'(k) = \rho + \delta - \frac{u''(c)}{u'(c)}\frac{dc}{dt}\cdot$$

This is our final equation for the optimal path of capital accumulation. It asserts that along such a path the rate of consumption at each moment must be chosen so that the marginal productivity of capital is the sum of three components:

(1) ρ, the social rate of time preference,
(2) δ, the rate of physical deterioration of capital, and
(3) the rather formidable looking third term which, however, is simply the percentage rate at which the psychic cost of saving diminishes through time. This can be seen by noting that the psychic cost of saving at any time is $u'(c)$, its time rate of change is $u''(c)dc/dt$, and its percentage time rate of change is the negative of the third term in the sum.

In other words, along the optimum path of accumulation the marginal contribution of a unit of capital to output during any short interval of time must be just sufficient to cover the three components of the social cost of possessing that unit of capital, namely, the social rate of time-preference, the rate of physical deterioration of capital, and the additional psychic cost of saving a unit at the beginning of the interval rather than at the end. All of these are expressed as percents per unit of time, which is also the dimension of the marginal productivity of capital.

The evolution of this economy along its optimal path of development can be visualized most readily by drawing a phase diagram as shown in Figure 1. We have found that the rates of change of k and c can be written:

$$\dot{k} = f(k) - (n+\delta)k - c, \tag{9}$$

$$\dot{c} = \frac{u'(c)}{u''(c)}[\rho + \delta - f'(k)].$$

Thus $\dot{k}=0$ whenever c and k satisfy the equation

$$c = f(k) - (n+\delta)k.$$

In Figure 1, k is plotted horizontally and c vertically. The curve labelled $\dot{k}=0$ shows the combinations of c and k that satisfy this equation. It has the shape drawn because of the conventional assumptions that the marginal productivity of capital is positive but diminishing (i.e., $f'(k)>0$,

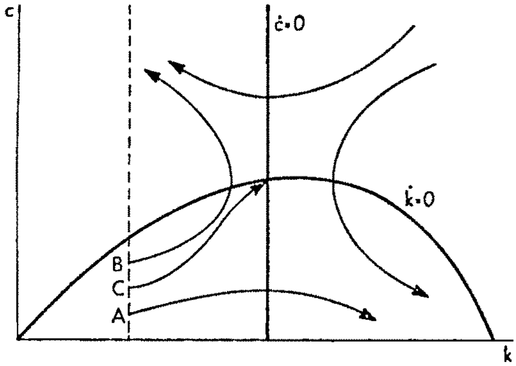


FIGURE 1

$f''(k)<0$), and the very plausible assumption that for very low levels of capital per worker, $f'(k)>n+\delta$. We also assume that no output is possible without some capital, i.e., $f(0)=0$. If consumption per capita is less than the rate on the locus just described, capital per capita increases ($\dot{k}>0$). Above the locus $\dot{k}<0$.

Similarly, consumption per capita is unchanging ($\dot{c}=0$) if

$$f'(k) = \rho + \delta.$$

The vertical line in Figure 1, labelled $\dot{c}=0$, is drawn at this level of k. If we accept the usual assumptions of positive but diminishing marginal utility $u'(c)>0$, $u''(c)<0$. Then $\dot{c}>0$, i.e., per capita consumption grows, to the left of this line. The reason is that with low levels of capital per capita the amount of depreciation is small and the amount of capital needed to equip the increment in population with the current level of capital per capita is also small.

These considerations enable us to depict qualitatively the laws of motion of the system. Imagine an initial low level of capital per capita, represented by the dashed vertical in the diagram. The entire evolution of the system is determined by the choice of the initial level of per capita consumption. If a low initial level is chosen, such as at point A in the figure, both consumption and capital per capita will increase for some time, following the curved arrow that emanates from point A. But when the level of capital per capita reaches the critical level, consumption per capita will start to fall though capital per capita will continue to increase. This is a policy of initial generosity in consumption followed by increasing abstemiousness intended, presumably to attain some desired ultimate level of capital per capita.

Similarly, the path emanating from point B represents a policy of continually increasing consumption per capita, with capital initially being accumulated and eventually being consumed. The other paths drawn have similar interpretations.

The path originating at point C is of particular interest. It leads to the intersection of the two critical loci, the steady state of the system in which neither per capita consumption nor per capita income changes. Once at this point all the absolute values grow exponentially at the common rate n.

It is now seen that if the initial capital per capita is given, the entire course of the economy is determined by the choice of the initial level of per capita consumption. This choice determines, among other things, the amount of capital per capita at any specified date.[13] If the conditions of the problem prescribe a particular amount of capital at some date, the initial c must be the one with a path that leads to the specified point. If there is no such prescription for capital accumulation, the initial c will be the one that causes the capital stock to be exhausted at the terminal date under consideration. And if there is no terminal date (i.e., $T=\infty$) the problem becomes much trickier mathematically and, indeed, the theory of optimization with an infinite time horizon is not yet completely established. But, in this simple case, we can see that the only

[13] The position of the economy at particular dates cannot be read off the phase diagram.

possible solution is the path that originates at point C and terminates at the point where $\dot{c}=\dot{k}=0$. For, the figure shows that all other paths that satisfy the optimizing conditions lead eventually to situations in which either c or k is negative. Since such paths cannot be realized, the only feasible optimizing path is the one that approaches $\dot{c}=\dot{k}=0$.

This result is quite characteristic of infinite horizon problems: the optimal growth paths, under many conditions, approach the situation in which consumption and the capital stock grow exponentially at a rate determined by the rate of population growth and the rate of technical progress (here assumed zero), just as in this case.

For finite horizon problems, it can be shown that the more remote the terminal date considered, the closer the path will come to the steady state position ($\dot{c}=\dot{k}=0$) before veering away to either high consumption or high capital accumulation as the case may be. This is a version of the turnpike theorem.

V. *Derivation via Finite Maximizing*

Those who distrust clever, intuitive arguments, as I do, may find some comfort in seeing the same results deduced from the more familiar method of maximizing subject to a finite number of constraints. Let us suppose that the entire period of T months is divided into n subperiods of m months each. $u(x_t, k_t, t)$ then denotes the rate at which profits are being earned or other benefits derived during the t-th subperiod, with x_t being the value of the decision variable during that subperiod, and k_t the value of the state variable at its beginning. Since the subperiod is m months long, the total profit earned is $u(x_t, k_t, t)\, m$.

The rate of change of the state variable during the t-th subperiod is $f(x_t, k_t, t)$. Then the values of the state variable at the beginnings of successive subperiods are connected by the equation

$$k_{t+1} = k_t + f(x_t, k_t, t)m. \tag{12}$$

Finally, the finite version of our problem is to choose $2n$ values, x_t, k_t so as to maximize the total profit over the entire period,

$$\sum_{t=1}^{n} u(x_t, k_t, t)m$$

subject to the n constraints (12), and to any boundary conditions that may apply. To be specific, suppose that initial and terminal values for the state variable are preassigned. These give rise to the side conditions

$$k_1 = K_0$$
$$k_{n+1} = K_T.$$

This problem is solved by setting up the Lagrangean function

$$\begin{aligned} L = &\sum_{t=1}^{n} u(x_t, k_t, t)m \\ &+ \sum_{1}^{n} \lambda_t[k_t + f(x_t, k_t, t)m - k_{t+1}] \\ &+ \lambda_0[K_0 - k_1] + \mu[k_{n+1} - K_T] \end{aligned}$$

and setting each of its partial derivatives equal to zero. The Greek symbols in this formula are the Lagrange multipliers, one for each constraint. We shall interpret them after we have completed our calculations.

The same Hamiltonian expression that we encountered before is beginning to emerge, so it is convenient to write

$$H(x_t, k_t, t) = u(x_t, k_t, t) + \lambda_t f(x_t, k_t, t)$$

and

$$\begin{aligned} L = m &\sum_{1}^{n} H(x_t, k_t, t) + \sum_{1}^{n} \lambda_t(k_t - k_{t+1}) \\ &+ \lambda_0(K_0 - k_1) + \mu(k_{n+1} - K_T). \end{aligned}$$

Now differentiate and equate derivatives to zero:

$$\frac{\partial L}{\partial x_t} = m \frac{\partial}{\partial x_t} H(x_t, k_t, \mathrm{t}) \tag{13}$$
$$= [u_1(x_t, k_t, \mathrm{t}) + \lambda_t f_1(x_t, k_t, \mathrm{t})]m = 0$$
$$\text{for } \mathrm{t} = 1, \ldots, n,$$

which is analogous to equation (5). And

$$\frac{\partial L}{\partial k_t} = m \frac{\partial}{\partial k_t} H(x_t, k_t, \mathrm{t}) + \lambda_t - \lambda_{t-1} = 0$$

or

$$-\frac{\lambda_t - \lambda_{t-1}}{m} = u_2(x_t, k_t, t) \tag{14}$$
$$+ \lambda_t f_2(x_t, k_t, \mathrm{t}), \text{ for } \mathrm{t} = 1, \ldots, n,$$

which is the discrete analog of equation (6).

Finally

$$\frac{\partial L}{\partial k_{n+1}} = -\lambda_n + \mu = 0.$$

Thus $\mu = \lambda_n$ and can be forgotten.

These equations are applicable to problems in which time is regarded as a discrete variable. The Lagrange multipliers have their usual interpretation. In particular, λ_t is the amount by which the maximum attainable value of $\sum u(x_t, k_t, \mathrm{t})m$ would be increased if an additional unit of capital were to become available by magic at the end of the t-th period. In other words, λ_t is the marginal value of capital on hand at date mt.

The maximizing conditions found previously should be the limit of these equations as m approaches zero and n approaches infinity, and they are. To show this, we have to revise our notation slightly. The subscripted variables now denote the values that the variables have in the t-th period. When m changes, the dates included in the t-th period change also. So we need symbols for the values of the variables at fixed dates. To this end, let τ denote any date and $x(\tau)$, for example, the value of x at that date. The connection between x_t and $x(\tau)$ is easy. Any date τ is in the subperiod numbered t where t is given by

$$\mathrm{t} = 1 + [\tau/m].$$

In this formula, [ ] is an old-fashioned notation meaning "integral part of." For example: [3.14159]=3. Then $x(\tau)$ is defined by

$$x(\tau) = x_{1+[\tau/m]},$$

and similarly for the other variables. Equations (13) and (14) can now be written in terms of τ:

$$u_1[x(\tau), k(\tau), \tau] \tag{15}$$
$$+ \lambda(\tau) f_1[x(\tau), k(\tau), \tau] = 0,$$

$$-\frac{\lambda(\tau) - \lambda(\tau - m)}{m} = u_2[x(\tau), k(\tau), \tau] \tag{16}$$
$$+ \lambda(\tau) f_2[x(\tau), k(\tau), \tau].$$

Notice in equation (16) that λ_{t-1} has been replaced by $\lambda(\tau - m)$, reflecting that the beginnings of the intervals are m months apart.

Equation (15) is identical with equation (II). As m approaches zero, the left-hand side of equation (16) approaches $-\dot{\lambda}(\tau)$, taking for granted that it approaches a limit and applying the definition of the derivative. The whole equation, therefore, approaches equation (III). Equation (I) is similarly and obviously the limiting form of equation (12).

Thus the basic equations of the maximum principle are seen to be the limiting forms of the ordinary first-order necessary conditions for a maximum applied to the same problem, and the auxiliary variables of the maximum principle are the limiting values of the Lagrange multipliers.

VI. *Qualifications and Extensions*

This entire development has been exceedingly informal, to put it kindly. The calculus of variations is a difficult and

delicate subject, so that a choice always has to be made between stating a proposition correctly, with all the qualifications that it deserves, and stating it forceably and clearly so that the essential idea can be grasped at a glance. The more intelligible alternative has been chosen throughout this paper since all the theorems have been stated and proved rigorously elsewhere in the literature.[13] This choice, as it happens, has especial drawbacks in the present context because much of the virtue of the maximum principle lies precisely in the qualifications that have been suppressed: it is valid under more general conditions than the classical methods that yield almost the same theorems.

As an example of the alternative mode of exposition, our main conclusions can be stated more formally and correctly as follows:[14]

THEOREM 1. Let it be desired to find a time-path of a control variable $x(t)$ so as to maximize the integral

$$\int_0^T u[k(t), x(t), t]dt$$

where

$$\frac{dk}{dt} = f[k(t), x(t), t],$$

where $k(0)$ is preassigned, and where it is required that $k(T) \geq \bar{K}$. It is assumed that the functions $u(k, x, t)$ and $f(k, x, t)$ are twice continuously differentiable with respect to k, differentiable with respect to x, and continuous with respect to t. Then if $x^*(t)$ is a solution to this problem, there exists an auxiliary variable $\lambda(t)$ such that:

(a) For each t, $x^*(t)$ maximizes $H[k(t), x(t), \lambda(t), t]$ where $H(k, x, \lambda, t) = u(k, x, t) + \lambda f(k, x, t)$;

(b) $\lambda(t)$ satisfies

$$\frac{d\lambda}{dt} = -\frac{\partial H}{\partial x}$$

evaluated at $k = k(t)$, $x = x^*(t)$, $\lambda = \lambda(t)$; and

(c) $k(T) \geq \bar{K}$, $\lambda(T) \geq 0$, $\lambda(t)[k(T) - \bar{K}] = 0$.

This theorem applies to the type of problem that we have been considering, with the useful elaboration that a lower limit has been imposed on the terminal value of the state variable, k. Part (c) of the conclusion, called the transversality condition, arises from this added requirement. It asserts that the terminal value of the auxiliary variable cannot be negative and that it will be zero if, at the end of the optimal path, $k(T)$ exceeds the required value.

The principle difference between this formal statement and our previous conclusions lies in conclusion (a) of the Theorem. The assertion that the Hamiltonian function, H, is maximized at each instant of time is not the same as the assertion that its partial derivatives vanish, made in our equations (II) and (II'). Equating partial derivatives to zero is neither necessary nor sufficient for maximization, though it is especially illuminating to economists, when it is appropriate, because conditions on partial derivatives translate readily into marginal equalities. There are three complications that can make the vanishing of partial derivatives an inadequate indication of the location of a maximum.

First, there are the so-called higher order conditions. First partial derivatives can vanish at a minimum or at a saddle-point as well as at a maximum. To guard against this possibility, second partial derivatives, and even higher ones, have to be taken into account.

Second, the vanishing of partial derivatives, even when higher order con-

[13] For example, in Arrow and Kurz [3] and Halkin [5].

[14] The given theorem is adapted from Arrow [2], Propositions 1 and 2. More elaborate theorems can be found in that source.

ditions are satisfied, establishes only a local maximum. It does not preclude that there may be some other value of the variables, a finite distance away, for which the function to be maximized has a still higher value. For reassurance on this point, one must inspect global rather than merely differential or local properties of the functions involved.

Finally, where the range of variation of the functions involved is limited in some manner, the maximum may be attained at a point where the partial derivatives do not vanish. This is a frequent occurence in economic applications, made familiar by linear programming. For example, it may be optimal for a firm with great growth possibilities to reduce its dividends to zero, though negative dividends are not permissible. In terms of our formulas this would be indicated by finding

$$\partial H/\partial x_t < 0 \quad \text{for all } x_t \geq 0,$$

where x_t denotes dividend payments per year at time t. H would be maximized by choosing $x_t=0$, its smallest permissible value, although the partial derivative does not vanish there.[15] This maximum could not be found by the ordinary methods of the calculus. Other methods are available, of course, for example those of mathematical programming. It is in just these circumstances that the maximum principle yields more elegant and manageable theorems than the older calculus of variations, which is more closely akin to the differential calculus.

For all these reasons, the fundamental condition for an optimal growth path is the maximization of $H(k, x, \lambda, t)$ at all moments of time, and the vanishing of $\partial H/\partial x$ is only an imperfectly reliable device for locating this maximum. It is, however, a very illuminating device and contains the conceptual essence of the matter, which is why we have concentrated on it.

[15] Technically this is called a "corner solution."

Throughout the discussion we have tried to be ambiguous about the exact nature of the time paths, $x(t)$ and $k(t)$. We have treated x and k as if they were one-dimensional variables, such as the quantity of capital or the rate of consumption. In many economic problems, however, there are several state variables and several choice variables. In such problems, it is profitable to think of $x(t)$, $k(t)$, their derivatives, and so on, as vectors. Then $\lambda(t)$ should also be regarded as a vector, with one component for each component of $k(t)$. When this viewpoint is taken, all our conclusions and the theorem still apply with scarcely a change in notation. That is why we were so ambiguous: it is easiest to think about ordinary numbers, but our conclusions and even most of our arguments are applicable when the variables are vectors.

The last remark raises some important new possibilities. Many economic problems concern the time paths of interconnected variables. For example, a problem may deal with the growth paths of consumption (c), investment (i), government expenditure (g), and income (y) in an economy. These four variables can be regarded as four components of a decision vector, x, connected by an income accounting identity $c(t)+i(t)+g(t)=y(t)$. Then the optimizing growth-path problem requires finding optimal growth paths for these four variables (and perhaps others) that satisfy the income accounting identity.

The new feature that we have encountered is the introduction of constraints or side conditions on the values of the decision variables. The same line of reasoning that we have been using applies, with the sole modification that when the function $V(k, x_t, t)$ is maximized, the vector x_t has to be chosen so as to satisfy all the side constraints. The algebra becomes somewhat more complicated but leads to conclusions like those discussed above and

with the same economic import. In 1968, Kenneth Arrow derived a lucid version of the formal statement of the theorem applicable to problems in which the decision variables are constrained. See [2, Proposition 3, p. 90]. Of course, this argument, too, presumes that circumstances are such that the proper partial derivatives vanish at the maximum.

REFERENCES

[1] K. J. ARROW, "Discounting and Public Investment Criteria," in A. V. Kneese and S. C. Smith, eds., *Water Research*, Washington 1966, pp. 13–32.

[2] ———, "Applications of Control Theory to Economic Growth," American Mathematical Society, *Mathematics of the Decision Sciences, Part 2*. Providence 1968, pp. 85–119.

[3] ———, AND M. KURZ, *Public Investment, the Rate of Return, and Optimal Fiscal Policy*. Stanford University Institute for Mathematical Studies in the Social Sciences, 1968.

[4] R. BELLMAN, *Dynamic Programming*. Princeton 1957.

[5] H. HALKIN, "On the Necessary Condition for Optimal Control of Nonlinear Systems," *Journal D'Analyse Mathématique*, 1964, *12*, 1–82.

[6] H. HOTELLING, "A General Mathematical Theory of Depreciation," *J. Amer. Statist. Ass.*, Sept. 1925, *20*, 340–53.

[7] L. S. PONTRYAGIN, V. G. BOLTYANSKII, R. V. GAMKRELIDZE, AND E. F. MISHCHENKO, *The Mathematical Theory of Optimal Processes*, (tr. by K. N. Trirogoff). New York 1962.

[8] F. P. RAMSEY, "A Mathematical Theory of Saving," *Econ. J.* Dec. 1942, *38*, 543–59.

[8] J. ROBINSON, *The Accumulation of Capital*. Homewood, 1956.

Money Illusion and the Aggregate Consumption Function

By WILLIAM H. BRANSON AND ALVIN K. KLEVORICK*

A standard result of the theory of rational consumer behavior in a static monetary economy is that a consumer's demand functions for commodities are homogeneous of degree zero in prices, money income, and money wealth.[1] Don Patinkin has defined this condition as the absence of money illusion. People whose demands for commodities would be altered by an equiproportionate change in all prices, money income, and money wealth are said to suffer from money illusion.

Aggregating over all commodities purchased by the consumer, this standard theorem leads to the conclusion that an individual's total real consumption demand is homogeneous of degree zero in prices, money income, and money wealth. Finally, aggregating over all consumers, this result would imply that the economy's aggregate real consumption should be a function of aggregate real income and aggregate real wealth, but not the price level.

Most empirical studies of aggregate consumer behavior assume this absence of money illusion when specifying their consumption functions. But the world in which consumers make their decisions and take their actions is quite different from the static model of traditional consumer theory where rationality and perfect information always prevail. First, the world we observe is a dynamic one. It is also one in which irrationality may exist in the short run and in which there are difficulties associated with the collection and interpretation of reliable information. Hence, while in the long run we might expect to find people free of money illusion, it is not so clear that in the short run we should expect to find consumers' total real consumption demand homogeneous of degree zero in prices, money income, and money wealth.

When one takes into account the lags that necessarily exist in processing price information in the real world, the case for price-level misperception by consumers becomes strong and the a priori case for the existence of money illusion in consumers' short-run demand functions becomes more convincing.[2] If one were, for example, to estimate an equation which explained current real consumption as a function of

* The authors are associate professor of economics and public affairs at Princeton University and assistant professor of economics at Yale University, respectively. The research described in this paper was partially supported by grants from the National Science Foundation and the Mobil Foundation. Computation was done at the Princeton University Computer Center using the program "TSP" written by Robert Hall, with an Almon subroutine by Charles Bischoff. Research assistance was provided by Raymond Hill and Michael Murphy of Princeton University. The authors are indebted to Charles Bischoff, William Brainard, Stephen Goldfeld, David Grether, Edward Kane, Richard Nelson, Marc Nerlove, William Nordhaus, Albert Rees, Joseph Stiglitz, James Tobin, and Roger Waud for their helpful comments and criticisms. They are not responsible for any faults that remain in the final product.

[1] See Patinkin [18, pp. 17–23, 403–05, and *passim*] and Samuelson [20, Ch. V].

[2] On the other hand, consider two economies which are structurally identical except that at time, $t=0$, the price level in one economy is k times the price level in the other. Allow both systems to run for a long time, and observe the parts of their respective time paths very far out in the future. One might be surprised if after a period long enough for the transient effects of the initial conditions to be removed from the two systems, one found the two systems behaving very differently. But it should be clear, at this point, that we are analyzing only the short-run relationship between consumption and income, wealth, and prices.

only current money income, money wealth, and prices, one would not be very surprised to find an inhomogeneity with respect to the current price variable. In fact, a crude test of the model presented in Section I, using only current values of all variables, yielded just such an inhomogeneity.[3]

We must then address ourselves to a basic question. If one estimates a short-run consumption function carefully, taking account of distributed-lag adjustments, simultaneous-equation relationships and the like, will the resulting short-run relationship show that money illusion is present? And if the price level does appear in the estimated equation in the form of a money-illusion effect, exactly how much of this money illusion is there?

The usual assumption that consumers are free of money illusion has not been subjected to systematic testing.[4] As Patinkin concluded at the end of his note on "Empirical Investigations of the Real-Balance Effect," "There are other basic questions which have not been dealt with here. Thus it would be desirable to carry out a direct test of the hypothesis that consumers are free of money illusion. . . ." [18, p. 664]

Only careful specification and examination of a per capita real consumption function can provide further insight and enable a more reasoned judgment to be made about the effect of the overall price level on consumer decisions and, in particular, about money illusion's presence. In Section I, we formulate a consumption function, in the framework of the Ando-Modigliani-Brumberg "life-cycle" hypothesis, that allows the price level to play an independent role in determining the level of per capita real consumption, and we discuss the alternative interpretations of this role. Fitting this function to *U.S.* quarterly data from 1955 I–1965 IV in Section II, we find that the general price level does, indeed, play a significant role in determining the level of per capita real consumption. Our results are shown to be consistent with a model which embodies a money-illusion effect via a distributed-lag adjustment to the price level or with a model which has the money-illusion effect combined with a price-expectations mechanism. But the results are not consistent with a model which hypothesizes the complete absence of money illusion. We conclude, then, that consumers do suffer from some degree of money illusion.

Section III investigates the degree to which our conclusions themselves might be illusory because of statistical problems. Our consideration of these possible statistical pitfalls leads to the conclusion that a significant and substantial degree of money illusion does exist in the *U.S.* consumption function. The paper concludes with a brief indication of the further questions our study's results would suggest should be examined.

[3] We estimated the consumption functions

$$\left(\frac{C}{P}\right)_t = \beta_0 + \beta_1\left(\frac{Y}{P^\alpha}\right)_t + \beta_2\left(\frac{W}{P^\alpha}\right)_t,$$

and

$$\left(\frac{C}{P}\right)_t = \beta_0 + \beta_1\left(\frac{Y}{P^\alpha}\right)_t + \beta_2\left(\frac{W}{P^\alpha}\right) + \beta_3\left(\frac{Y}{P^\alpha}\right)_t\left(\frac{W}{P^\alpha}\right)_t,$$

where C is per capita consumption, Y is per capita labor income, W is per capita consumer net worth, P is the price level of consumer goods, and α is a money-illusion parameter. (Clearly, $\alpha=1$ implies the absence of money illusion, while $\alpha=0$ implies extreme money illusion in the Patinkin sense. See [10].) Our results indicated that the maxima of the likelihood functions for the two equations were quite far from the point $\alpha=1$, being in fact much closer to $\alpha=0.00$ than to $\alpha=1$.

[4] For example, consumption functions have generally been estimated using *either* money *or* real values of the variables. Rarely is any attempt even made to compare the results obtained using the same specification except for the substitution of real for money values or vice versa. Two exceptions are studies by R. Ferber [8] and J. J. Arena [3], [4]. Unfortunately they did not yield any significant conclusions about money illusion's existence.

I. *The Specification of a Money-Illusion Consumption Function*

The life-cycle analysis of rational consumption behavior begins with a consumer who maximizes his utility function subject to the constraint imposed by his resources.[5] "As a result of this maximization, the current consumption of the individual can be expressed as a function of his resources and the rate of return on capital with parameters depending on age." [2, p. 56] Aggregating across all consumers and then dividing by population, one arrives at an Ando-Modigliani-Brumberg life-cycle per capita consumption function:

$$\frac{C}{P} = f\left(\frac{Y}{P}, \frac{W}{P}\right), \tag{1}$$

where C is per capita consumption, Y is per capita nonproperty income, W is per capita consumer net worth, and P is the price level of consumer goods.

Without more detailed assumptions about individual utility functions and distributional effects of changes in Y and W, the only meaningful restriction we can impose on (1) is that it is homogeneous of degree zero in Y, W, and P.[6] Therefore we can specify the aggregate consumption function (1) in multiplicative form as:

$$\left(\frac{C}{P}\right)_t = b_0' \left(\frac{Y}{P}\right)_t^{b_1} \left(\frac{W}{P}\right)_t^{b_2}. \tag{2}$$

If money illusion exists, however, the price level should have an independent effect on the level of C/P. Letting c denote real consumption, y denote real net labor income, and w denote real consumer net worth, all on a per capita basis, the consumption function in the presence of money illusion then takes the form

$$c_t = b_0' (y_t)^{b_1}(w_t)^{b_2}(P_t)^{b_3}. \tag{3}$$

This log-linear specification, (3), can be justified by noting that beyond requiring that the form chosen possess a certain set of characteristics, the choice of a particular form for a consumption function is an arbitrary process. One should, for example, ensure the proper signs and magnitudes of relevant partial derivatives—marginal propensities to consume out of labor income and out of wealth that are positive and less than unity. The signs and magnitudes of these marginal propensities can only be checked *ex post*, although we can assure the reader that the estimates we obtain do meet these prior specifications.

In logarithmic form, the consumption function in (3) can be rewritten as

$$ln\, c_t = b_0 + b_1\, ln\, y_t + b_2\, ln\, w_t + b_3\, ln\, P_t. \tag{4}$$

With the consumption function written in this form, if there is no money illusion in the Patinkin sense, we have $b_3 = 0$, and real consumption depends only on real net labor income and real consumer net worth. On the other hand, if there is money illusion in the Patinkin sense—a proportional increase in money income, money wealth, and the price level leads to an increase in the level of real consumption—we have $b_3 > 0$.

There is, however, no reason to expect consumers to react instantaneously, that is, within one quarter, to changes in real income, real wealth, or the price level. It is much more plausible to suppose that consumers react with a lag to changes in these independent variables. Alternatively, one might think it plausible that in making their real consumption decisions in a particular quarter, consumers consider an average of recent experience with regard to the consumption determining variables. Thus, we will rewrite (4) to allow for the possibility of distributed-lag adjustments

[5] See Modigliani and Brumberg [15], [16] and Ando and Modigliani [2] for the development of the life cycle hypothesis of saving.

[6] [2, p. 58]. Ando and Modigliani present a much more careful derivation, based upon carefully enunciated assumptions, of their linear "life-cycle" hypothesis model. See pp. 56–62 therein.

to income, wealth, and prices. Using this distributed-lag model, the basic equation to be estimated is the following:

$$\ln c_t = \beta_0 + \sum_{i=0}^{I} \gamma_i \ln y_{t-i} + \sum_{j=0}^{J} \delta_j \ln w_{t-j} + \sum_{k=0}^{K} \eta_k \ln P_{t-k} + \epsilon_t. \tag{5}$$

Before going on to estimate the money-illusion consumption function (5), it will be useful to discuss the interpretation of the price term,

$$\sum_{k=0}^{K} \eta_k \ln P_{t-k},$$

and to make clear its role in testing the no money-illusion hypothesis. There are at least three ways in which a price-level effect might make its appearance in our basic equation. Since money illusion represents only one such effect and since these three possibilities are not mutually exclusive, it is important to show how one can determine whether money illusion is present in consumer behavior. It will be helpful in examining these three effects to write the basic equation in multiplicative form, equation (6). We see that the first type of price effect possible is the case of pure money illusion in the Patinkin sense. Instead of basing their consumption decisions on real income and real wealth, consumers modify the deflating factors of income,

$$\prod_{i=0}^{I} P_{t-i}^{-\gamma_i}, \quad \text{and wealth,} \quad \prod_{j=0}^{J} P_{t-j}^{-\delta_j},$$

by multiplying their product by

$$\prod_{k=0}^{K} P_{t-k}^{\eta_k} \neq 1.$$

When prices, money income, and money wealth all increase proportionately, consumers notice the income and wealth increases more than they do the price level rise, and increase their real consumption. Hence, in the case of *pure money illusion we would have*

$$\sum_{k=0}^{K} \eta_k > 0:$$

consumers exhibit money illusion via a distributed-lag adjustment to the price level.

Suppose, in contrast, that consumers do not suffer from money illusion, but that there is a price-expectations mechanism at work. That is, the real consumption function takes the form

$$c_t = g(y_{t-i}, w_{t-j}, P_t^e), \tag{7}$$

where y_{t-i} is a vector of recent real income experience, w_{t-j} is a vector of recent real wealth experience, and P_t^e is a vector of the consumers' expectation of future price levels. The hypothesized behavior lying behind such a function is that if consumers expect prices to rise in the future, they will restructure the time pattern of their consumption by moving consumption from the future toward the present. Then, if their expectations are realized, they will reduce their consumption in the future.[7]

Price expectations, the vector P_t^e, could be formed in several different ways. There are two formation processes on which we will focus here in order to see how a money-illusion effect can be identified. They are expectations derived on the basis of recent price *level* experience,

[7] See Power [19] for a further discussion of the role of the intertemporal substitution effect, price expectations, and the real-balance effect.

$$c_t = e^{\beta_0}\left[\prod_{i=0}^{I}\left(\frac{Y}{P}\right)_{t-i}^{\gamma_i}\right]\left[\prod_{j=0}^{J}\left(\frac{W}{P}\right)_{t-j}^{\delta_j}\right]\left[\prod_{k=0}^{K} P_{t-k}^{\eta_k}\right] e^{\epsilon_t}. \tag{6}$$

$$P_t^e = h_1(P_{t-k}), \tag{8}$$

where P_{t-k} is the vector of recent price experience, and expectations based on recent observations of the *rate* of change of the price level,

$$P_t^e = h_2(\Delta P_{t-k}), \tag{9}$$

where ΔP_{t-k} is the vector of recent price change experience.

Considering the level-based expectations mechanism first, in terms of our log-linear consumption function P_t^e would have to enter into the consumption decision in the form of a product of the P_{t-k}'s,

$$\prod_{k=0}^{K} P_{t-k}^{\eta_k},$$

with the weights summing to zero. The weights must sum to zero because of the purely allocative role of the price-anticipations mechanism and because in a steady state with $P_{t-1}=P_t$ for all t, P_t^e must have no effect on the consumption decision as represented in (7): prices will not be changing. In the case of a *pure level-based price-expectations mechanism and no money illusion, we would have*

$$\sum_{k=0}^{K} \eta_k = 0.$$

On the other hand, if consumers' price expectations were formed on the basis of recent price-level changes as in (9) (referred to here as change-based expectations), a log-linear consumption function would take the form of equation (10), where the price ratios represent the rates of change indicated in equation (9). In this case, the purely allocative role of expectations about future prices implies that the sum of the θ_k parameters should be zero. Note, moreover, that the consumption function in equation (10) is consistent with the behavior we should expect in a steady state. If there exists a steady state, so that prices do not change from one period to the next, the P_t^e argument should disappear from the function in (7). This is precisely what equation (10) indicates would happen, since with $P_{t-1}=P_t$ for all t, each P_{t-k}/P_{t-k-1} ratio would equal unity and we would have

$$c_t = e^{\beta_0}\left[\prod_{i=0}^{I} y_{t-i}^{\gamma_i}\right]\left[\prod_{j=0}^{J} w_{t-j}^{\delta_j}\right]e^{\epsilon_t}.$$

The question remains as to how we should expect the estimated price coefficients in equation (5) to appear if there is no money illusion but the change-based price-expectations mechanism just described exists. The answer is really quite simple. Writing (10) in logarithmic form, we have

$$\begin{aligned} \ln c_t = {} & \beta_0 + \sum_{i=0}^{I} \gamma_i \ln y_{t-i} \\ & + \sum_{j=0}^{J} \delta_j \ln w_{t-j} \\ & + \sum_{k=0}^{K-1} \theta_k(\ln P_{t-k} - \ln P_{t-k-1}) \\ & + \epsilon_t. \end{aligned} \tag{11}$$

But the price term can be written as

$$\begin{aligned} & \sum_{k=0}^{K-1} \theta_k(\ln P_{t-k} - \ln P_{t-k-1}) \\ & = \theta_0 \ln P_t + \sum_{k=1}^{K-1} (\theta_k - \theta_{k-1}) \ln P_{t-k} \\ & \quad - \theta_{K-1} \ln P_{t-K}. \end{aligned} \tag{12}$$

Therefore, if we estimate (5) and the true

$$c_t = e^{\beta_0}\left[\prod_{i=0}^{I} y_{t-i}^{\gamma_i}\right]\left[\prod_{j=0}^{J} w_{t-j}^{\delta_j}\right]\left[\prod_{k=0}^{K-1}\left(\frac{P_{t-k}}{P_{t-k-1}}\right)^{\theta_k}\right]e^{\epsilon_t}. \tag{10}$$

model is one of pure change-based price-expectations and absence of money illusion, that is (10), we would have

$$\sum_{k=0}^{K} \eta_k = \theta_0 + \sum_{k=1}^{K-1} (\theta_k - \theta_{k-1}) - \theta_{K-1} = 0. \tag{13}$$

Thus, in the case of a *pure change-based price-expectations model and no money illusion, we would have*

$$\sum_{k=0}^{K} \eta_k = 0.$$

This discussion shows that estimation of the basic equation (5) will yield an unambiguous test of the no money-illusion hypothesis. If money illusion is present to some degree, we should find

$$\sum_{k=0}^{K} \eta_k$$

positive and significantly different from zero. If, on the other hand, the true model is one in which money illusion is absent we should find that

$$\sum_{k=0}^{K} \eta_k$$

is not significantly different from zero. It should be stressed that we cannot distinguish between (a) a model in which money illusion is present but there is neither a level-based nor a change-based price-expectations mechanism at work, and (b) a model in which money illusion is present and such price-expectations mechanisms are also operative. Ideally, one would like to distinguish among the following alternative deviations from the standard static model of consumer behavior: (1) the existence of only pure expectations mechanisms, (2) the presence of only short-run money illusion, and (3) the existence of a combination of some expectations mechanism and money illusion. The statistical estimation and tests to which we now turn enable us to distinguish alternative (1) from alternatives (2) and (3), but we are unable to distinguish between the latter two possibilities.

II. *Estimation of the Money-Illusion Consumption Function*

The money-illusion consumption function, equation (5), will now be estimated. Section III will then consider various potential problems involved in the estimation procedure, for example, common trends, simultaneity, and so on.

The data used are *U.S.* quarterly series on real consumption per capita, c, real net labor income per capita, y, real consumer net worth per capita, w, and the price level, P, for the period 1955-I to 1965-IV. These data are described in more detail in the Appendix.

The Consumer Price Index (*CPI*) (1958 = 100) was chosen as the price variable, P, since it represents the set of prices most relevant to the consumer's buying decision. Use of the principal alternative price variable, the consumption deflator, would create a statistical difficulty since the current value of the price term would be the deflator of the dependent variable, consumption. The denominator of the left-hand side of the regression equation would then appear in the numerator of the right-hand side of that equation. This would cause the coefficient of the current price variable to be negative, and, since the price series is serially correlated, it would also reduce the coefficients of other recent values of the price variable, all due to a statistical aberration. The *CPI*, of course, has its drawbacks as well.[8]

[8] For example, being a Laspeyres price index, it is subject to the customary catalogue of criticisms that can be levelled at such a fixed-weight-base index. In particular, it does not take accurate account of changes in the market basket that result from changes in the relative prices of commodities, and it does not provide an adequate means for coping with the introduction of

Given the different shortcomings of the alternative price indices, what is crucial is that our results concerning the significance or insignificance of the price term in the consumption function should be independent of the choice of index. The *CPI* will be used as the price variable on the right-hand side of the basic equation (5), but we will also show the result of substituting the consumption deflator and the deflator for personal consumption expenditures—indexes in which price relatives receive shifting weights in proportion to the expenditures incurred each year for the goods and services they represent. We can assure the reader that this substitution does not affect the qualitative results of our study.

The data on consumption, income, wealth, and prices will be used in estimating the basic distributed-lag consumption function equation (5), shown previously. An I quarter lag distribution assigns nonzero values to the coefficients of the variable lagged $0, 1, \ldots I-1$ quarters and a zero value to the coefficients of the variable lagged $I, I+1, \ldots$ quarters. Since the basic equation is linear in natural logarithms the estimated coefficients are, of course, estimates of the elasticity of real consumption demand with respect to changes in y, w, and P.

Each of the independent variables is entered in the form of a distributed lag with current real consumption per capita dependent on current and past values of the independent variables. The distributions of the coefficients of these lagged independent variables, which show the time shape of response of c to changes in y, w, and P, are estimated using the flexible Almon interpolation technique.[9] This method takes the lagged values of each of the independent variables as a set and estimates a separate smooth distribution of coefficients for each variable, subject only to the constraint that the coefficients be interpolated from Lagrange polynomials of a given degree.

Since two critical values might be needed to capture the lag distribution on the price terms if the purely allocative level-based expectations hypothesis discussed in Section I were correct, third-degree polynomials are used in estimating the coefficients of the price terms in (5). We will also use third-degree polynomials in estimating the distributed lags of income and wealth in (5). The freedom this accords to the shape of the distributions of the income, wealth, and price coefficients will ensure that our results do not come from the imposition of monotonic lag distributions on the coefficients.[10]

Since changing the lag length on one variable in (5) will usually affect the coefficients of all terms, searching for the optimal lag length on all three variables is a fairly complex procedure. In addition, there can sometimes be a conflict between criteria for determining the best lag length: overall goodness of fit, significance of the last coefficient, shape of the lag distribu-

new commodities or quality changes. For an excellent discussion of the problems associated with the *CPI* as a price index, see National Bureau of Economic Research [22, especially Ch. IV].

[9] See Almon [1] for the basic theory. Almon, Bischoff [5] and Modigliani and Sutch [17] have all used the Almon technique extensively.

[10] The Almon technique permits the user to constrain the coefficient of the value of the independent variable one period forward, for example, $\ln y_{t+1}$, to equal zero, and/or to constrain the distribution of coefficients to taper off gradually to zero at the far end, e.g., as i approaches I, by setting the last coefficient equal to zero. Since we do not want to exclude the possibility of convex monotonically decreasing lag distributions, we will not apply the zero constraint to the coefficients of the one-period ahead values of the independent variables. But since we expect the coefficient distributions, whatever their shapes, to approach zero gradually rather than abruptly, as the relevant variable values recede into the past, we will constrain the distributions to taper off gradually. This constraint will smooth the distribution somewhat and it should yield the same length "best" lag distribution as would unconstrained estimation. Our estimated equation is virtually the same in both the constrained and unconstrained versions, as a glance at fn. 12 will confirm.

tion, and so on. We began by setting lag lengths I, J, and K in (5) all at four quarters, and then experimented with changes in those lengths.

With $I=J=K=4$ in the initial estimate of (5), the price coefficients were all positive with a significant sum. Only the current wealth coefficient was at all significant, and the income lag was obviously too short—the coefficient of y_{t-3} was significantly positive. We therefore lengthened the lag distribution on income and shortened that on wealth until we reached the first equation shown in Table 1, in which $I=7$; $J=1$; $K=4$. Table 1 lists the coefficients of each equation horizontally, with the first number for each variable giving the lag length and the second the sum of the coefficients in the lag distribution for that variable, with the standard error of the sum in parentheses. Figure 1 shows the lag distribution of the coefficients of $\ln y$ and $\ln P$ in Table 1 equations 1-1 through 1-4.

The regressions that led to equation 1-1 from the initial 4-4-4 specification showed that while the current wealth term was highly significant in all cases, lagged wealth terms were uniformly insignificant, and always quite near zero, whenever the wealth lag was extended beyond one quarter. This might be expected because (a) the wealth series is highly autocorrelated, and (b) one might expect a weighted average of past labor incomes to be collinear with the wealth of very recent periods.

Equation 1-1 with the income lag at seven quarters and the price lag at four quarters shows that the coefficients of all three explanatory variables are highly significant. The income lag, in Figure 1, is positive and monotonically declining, while the price lag is positive in the shape of an inverted U. Extending the length of the price lag gave us equations 1-2 to 1-4 of Table 1. The effect of lengthening the price lag, as can be seen in Figure 1, was to change its shape from a significant inverted U to a significant monotonically declining distribution. But, as is shown in Table 1, the sum of the price coefficients rises only slightly, from 0.411 to 0.418,

TABLE 1—ESTIMATION OF THE BASIC MONEY-ILLUSION CONSUMPTION FUNCTION TEXT EQUATION (5)

Equation	Independent Variables									
	Constant	$\ln y$ I	$\ln y$ $\sum\gamma_i$	$\ln w$ J	$\ln w$ δ_0	$\ln P$ K	$\ln P$ $\sum\eta_k$	R^2	$SE\times10^2$	$D.W.$
1-1	−1.937 (0.115)	7	0.647 (0.042)	1	0.136 (0.036)	4	0.411 (0.036)	.9982	.3006	1.727
1-2	−1.945 (0.114)	7	0.658 (0.044)	1	0.129 (0.036)	5	0.415 (0.036)	.9982	.2980	1.756
1-3	−1.952 (0.113)	7	0.662 (0.044)	1	0.126 (0.033)	6	0.418 (0.036)	.9983	.2964	1.755
1-4	−1.953 (0.114)	7	0.661 (0.043)	1	0.127 (0.036)	7	0.418 (0.036)	.998¾	.2964	1.757
1-5	−1.941 (0.117)	8	0.659 (0.046)	1	0.132 (0.035)	7	0.413 (0.036)	.9982	.2990	1.746
1-6	−1.943 (0.118)	8	0.658 (0.046)	1	0.133 (0.035)	8	0.413 (0.036)	.9982	.2988	1.744

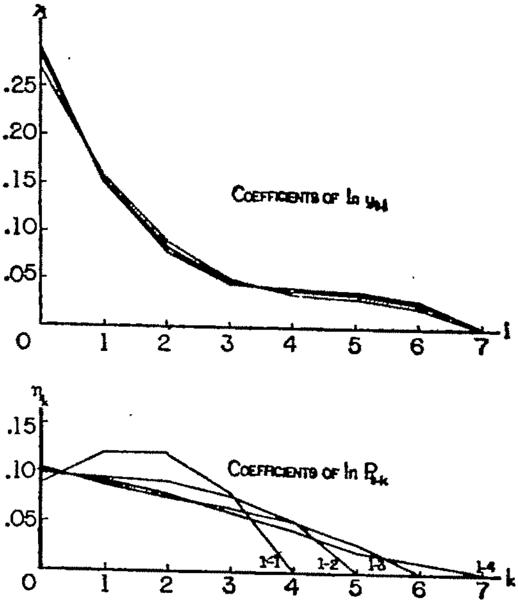


FIGURE 1. DISTRIBUTION OF COEFFICIENTS OF $ln\ y_{t-i}$ AND $ln\ P_{t-k}$ IN TABLE 1 EQUATIONS 1-1 TO 1-4

while the entire distribution becomes more significant. Lengthening the price lag from four to seven quarters leaves the income and wealth coefficients substantially unchanged. It does, however, reduce the standard error of the estimate of real consumption per capita from \$5.03 in 1-1 to \$4.99 in 1-4, compared with an average per capita consumption of \$1,837, and it raises the Durbin-Watson statistic from 1.73 to 1.76, both well above the range normally encountered in this type of time-series estimation when no lagged dependent variables are explicitly included.

When the lag on the wealth variable was extended in any of the Table 1 equations, the coefficients of all wealth terms but the current one were completely insignificant. The ratio of their sum to its standard error was less than the t-ratio of w_t alone in the unlagged version while the income and price coefficients and the equation statistics in these regressions were not significantly different from those of equation 1-4. This evidence led us to include only the current value of wealth in our equations.

Equation 1-4 of Table 1 has been chosen as our best estimate of the money-illusion consumption function. While it is only marginally superior to equations 1-1 to 1-3 in a statistical sense, it does have more significant coefficients in both the income and price lags, a lower standard error than that of 1-1 and 1-2, and a higher Durbin-Watson statistic. Also, if one believes that real consumption depends on present and lagged values of money income with each value deflated by a corresponding misperceived price level, then one would think that the price lag and income lag should be roughly the same length since the variable really moving consumption is incorrectly deflated income.[11]

Our choice of 1-4 as the estimate of the money-illusion consumption function is buttressed by the fact that when we extend the lags on income and price beyond seven quarters, the standard error rises again, the Durbin-Watson decreases, and the significance of both the income and price lags falls. Two examples of the effect of extending these lags appear in equations 1-5 and 1-6. Furthermore, equation 1-6 with an eight-period lag on both price and income is better than equation 1-5 which has a seven-period price lag and an eight-period income lag. This tends to support the belief that the price and income lag should be the same length.

The final equation for the money-illusion consumption function is, then,

$$(14) \qquad ln\ c_t = -\underset{(0.114)}{1.953} + \sum_{0}^{7} \gamma_i\ ln\ y_{t-i} +$$

[11] It is interesting to note that with monotonically declining lag distributions of equal length on both price and income, the money-illusion consumption function could reflect an adaptive adjustment process with real consumption following incorrectly deflated income in the manner suggested by Koyck [11]. See Griliches [9] for a review of the Koyck and similar models.

$$+ \underset{(0.036)}{0.127} \ln w_t + \sum_{0}^{7} \eta_k \ln P_{t-k}$$

$R^2 = .9984$; $S.E. = .002964$; $D.W. = 1.757$.

The standard error of predicted c implied by (14) is only \$4.99 compared with a mean c of \$1,837.

The coefficients of lagged per capita income, γ_i, and lagged price, η_k, are plotted in Figure 1 and listed in Table 2.[12] The coefficients of real net labor income lagged zero to six quarters sum to 0.661 in our final equation. This is the elasticity of per capita real consumption with respect to changes in per capita real net labor income. The implied marginal propensity to consume is .71 at 1965-IV levels of \$2,118 for per capita consumption and \$1,964 for per capita income. Similarly, the per capita real wealth coefficient and elasticity is 0.127, giving a marginal propensity to consume out of real net wealth of .024 at 1965-IV per capita wealth of \$11,160. These marginal coefficients can be compared with the Ando-Modigliani values of .70 and .06 on income and on wealth.

The more interesting coefficients are those of the current and lagged values of the *CPI*. While the exact shape of the price lag is not completely clear, as a glance at Figure 1 can show, in all the versions of the money-illusion consumption function shown in Table 1, the sum of the price coefficients is between 0.411 and 0.418—the difference of .007 is completely insignificant. Furthermore, the sum of the price coefficients is highly significant in all versions of the equation; $\sum \eta_k$ is never less than eleven times its standard error. Thus real consumption rises when the *CPI* rises, real income and wealth being held constant, with an elasticity of 0.418 in our final equation.[13] If the *CPI* rises by 1 percent (not percentage point) consumption rises by 0.418 percent, or, at 1965-IV levels, \$8.85 (in 1958 dollars).

This sensitivity of real consumption to the price level in our log-linear specification of the consumption function will, of course, lead to the conclusion that real consumption will exceed Gross National Product (*GNP*) if prices rise relative to real income

TABLE 2—COEFFICIENTS OF $\ln y_{t-i}$ AND $\ln P_{t-k}$ IN EQUATION 1-4 OF TABLE 1, TEXT EQUATION (14)

Lag (i, k)	Coefficient of $\ln y_{t-i}$	Coefficient of $\ln P_{t-k}$
0	0.274 (0.046)*	0.100 (0.082)
1	0.151 (0.015)	0.093 (0.029)
2	0.081 (0.021)	0.080 (0.047)
3	0.048 (0.018)	0.063 (0.039)
4	0.039 (0.013)	0.044 (0.023)
5	0.038 (0.019)	0.026 (0.029)
6	0.030 (0.019)	0.010 (0.032)
7	0	0
Sum	0.661	0.418
Standard error of Sum	(0.043)	(0.036)

* The numbers in parentheses are the standard errors of the coefficients.

[12] Reestimation of text equation (14) without the constraint that the distributions of the coefficients of $\ln y_{t-i}$ and $\ln P_{t-k}$ taper off to zero as i and k approach 7 does not significantly change the equation. The income lag in the unconstrained equation has a coefficient sum of 0.662 with a standard error of 0.045; the coefficient of $\ln y_{t-8}$ is -0.003 with a standard error of 0.058. The price lag has a coefficient sum of 0.431 with a standard error of 0.039; the coefficient of $\ln P_{t-8}$ is -0.045 with a standard error of 0.094. The coefficient of $\ln w_t$ is 0.115 with a standard error of 0.038. $R^2 = .9983$; S.E. $= .003001$; $D.W. = 1.783$.

[13] This positive value for $\partial c/\partial P$ represents the presence of money illusion in the sense of the traditional Patinkin experiment—double all prices, money income and money wealth and see if real consumption rises.

and real wealth for a long enough period of time. It should be recalled at this point, though, that the relationship (14) presents is a *short-run* consumption function, without any necessary long-run implications for rationality or money illusion.[14]

To ensure that our results concerning the significance of the price variable in determining real consumption per capita are not sensitive to the use of the *CPI* as the price variable, we reestimated equation (14) using first the consumption deflator and then the deflator for personal consumption expenditures as the price variable. The use of these deflators leads one to observe the statistical artifact mentioned at the beginning of this section. In both equations the distribution of the price coefficients had the shape of an inverted U with the coefficient of the current value of the price variable insignificantly negative and that of $ln\, P_{t-1}$ insignificantly positive. Our qualitative conclusions derived from (14) were not, however, affected by either of these reestimated equations.[15]

Finally, to test the sensitivity of our results to the particular income series used, we also reestimated (14) using disposable personal income per capita deflated by the consumption deflator in place of real net labor income per capita. As was the case with the price variable, our qualitative conclusions, particularly those concerning the role of prices in determining real consumption, were not sensitive to the choice of income series.[16]

The results presented in this section show that the price level has a significant, independent, positive effect on the level of per capita real consumption. When prices, money income, and money wealth rise in the same proportion, real consumption rises. We interpret this result as evidence that there exists a significant degree of money illusion in the economy in the short run.

As the discussion in Section I showed, the price level might also affect the level of real consumption through a price-expectations mechanism. But it was also shown in that section that the conclusive test for the presence or absence of money illusion in our specification of the consumption function was whether or not $\sum \eta_k$, the sum of the price coefficients, was positive and significantly different from zero. Since the estimates in this section lead to the conclusion that the sum of the price coefficients in (5) is significantly positive, our results are inconsistent with either a pure level-based or a pure change-based price-expectations hypothesis.

Further evidence supporting this conclusion can be obtained by looking at the effect of explicitly imposing the constraint that the sum of the price coefficients in (5) is zero. As the discussion in Section I showed, this restriction is imposed if one estimates equation (15) instead of (5):

$$ln\, c_t = \beta_0 + \sum_{0}^{I} \gamma_i\, ln\, y_{t-i} + \sum_{0}^{J} \delta_j\, ln\, w_{t-j} + \sum_{0}^{K-1} \theta_k\, ln \frac{P_{t-k}}{P_{t-k-1}} + \epsilon_t. \tag{15}$$

[14] As it stands, the objection that raising prices relative to income and wealth for a long enough time leads to absurd results simply says that if we extrapolate our results sufficiently far beyond the data the results become invalid. This is not a surprising result.

[15] The coefficient sum of the consumption deflator was 0.332 with a standard error of 0.042; the coefficient sum using the deflator for personal consumption expenditures was 0.366 with a standard error of 0.043. The income and wealth coefficients in both equations were not significantly different from those of (14). In both cases the standard error of estimate was slightly higher than that of (14) and the Durbin-Watson statistic was lower, suggesting that the *CPI* may well be, in fact, the correct price series to use.

[16] As expected, substituting real disposable personal income per capita for real net labor income per capita slightly raised the sum of the income coefficients to 0.678, and reduced the wealth coefficient to 0.082. It also reduced the sum of the *CPI* price coefficients to 0.338 with a standard error of 0.029.

Setting $I=7$, $J=1$, and $K\text{-}1=6$, the appropriate lag lengths determined by our earlier work, we obtained the following estimates:

β_0	-0.678 (0.066)	K-1	6
I	7	$\Sigma\theta_k$	-0.004 (0.643)
$\Sigma\gamma_i$	0.533 (0.092)	R^2	0.9918
J	1	$S.E.\times 10^2$	.6416
δ_0	0.453 (0.049)	$D.W.$	0.817

The standard error of this estimate of (15) is more than twice that of our final equation (14), and the Durbin-Watson statistic is only 0.82, strongly suggesting that the equation has been misspecified.

That this estimate of (15) is significantly inferior to the money-illusion consumption function, equation (14), can also be seen from the following analysis of variance. The sum of squared residuals in (14) is $.3163\times 10^{-3}$; that in the estimate of (15) is $.1482\times 10^{-2}$. With forty-four observations and eight regression variables in each equation, we have

$$F(1, 36) = \frac{1.4820 - .3163}{.3163/36} = 132.7$$

to test the significance of the effect of the added restriction that $\sum \eta_k$ is zero. Since $F(1, 36)=7.39$ at the 1 percent level, it is clear that constraining $\sum \eta_k$ to equal zero significantly worsens the explanatory power of the equation.[17]

This estimate of (15) and the corresponding F-test provide additional evidence that we are not observing simply a pure price-expectations mechanism. There may be a price-expectations mechanism at work in the determination of real consumption, but if there is, it is operating in conjunction with the existence of money illusion.

III. *Statistical Problems of Trend, Cycle, and Simultaneity*

This section reports briefly on several further tests of the money-illusion consumption function, equation (14), which were conducted to ensure that our results are not seriously affected by problems of time and timing: trend interrelationships among variables, cyclical factors in the economy, and simultaneous equations bias.

Trend Relationships Among Variables

In any time-series regression analysis there exists the possibility that a spurious fit may be obtained due to the fortuitous presence of trends in both dependent and independent variables. While the Durbin-Watson statistic of equation (14), 1.76, suggests that we have captured more than a trend relationship, we performed a direct test of the role of time trend in our results by regressing *ln c/ct* the natural log of consumption deviations from trend on similar transformations of the income, wealth and price variables.[18] This is, of course, equiva-

[17] A search for the "best fitting" estimate of (15) yielded an equation with an eleven-quarter income lag and a seven-quarter price lag. The estimate of (15) with $I=11$, $J=1$ and $K\text{-}1=7$ fits better than the one presented in the text but still has a standard error nearly twice that of any of the equations in Table 1, and a Durbin-Watson statistic less than unity. Its F-test performance was correspondingly poor as F (1, 36) was 104.9.

[18] We should note that if the *CPI* has a spurious upward trend due to its inadequacy in observing quality changes and the introduction of new products, and if this effect somehow underlies the results in Section II, then the sum of the coefficients of the price term in the deviations-from-trend equation should be insignificant. Also, if the changing age distribution of the population was increasing the aggregate propensity to consume, and if this movement was being picked up by the trend in our price variable, the sum of the price coefficients in the deviations-from-trend version should be insignificant. In fact, the percentage of all families with head aged 15–25 has been growing, but the average income of these families is only about 75 percent of the national average. Hence the effect of this population drift on the propensity to consume is ambiguous in any event. Branson and Thurow in an unpublished paper [7] have concluded that changing population distribution, in the period 1950–1970, has little effect on the aggregate propensity to consume.

lent to simply adding time to equation (14) as an independent variable.

Equation 3–1, Table 3, shows the result of reestimating the money-illusion consumption equation (14) using deviations from trend. The format of Table 3 is the same as that of Table 1 except that since lag lengths are fixed at $I=7$, $J=1$, $K=7$, they are not shown. This deviations-from-trend version of the consumption function explains 95 percent of the variance of the deviations of consumption from its logarithmic trend. All the independent variables are significant. In particular, the sum of the price coefficients is still significantly positive, 2.7 times its standard error, indicating that real consumption per capita is positively related to the price level both along trend and in deviations from trend.

Price Movements and the GNP Gap

Another potential role of the price level in determining consumption, besides the existence of money illusion or the existence of price-expectations mechanisms, could be the presumed correlation of price movements with employment and distributional factors in the business cycle. It might be possible, for example, that as aggregate demand rises relative to potential output and unemployment falls, prices rise. The falling unemployment rate could increase the income of low-income families with higher-than-average consumption propensities, shifting the per capita consumption function (of income and wealth alone) up. If this movement were generally associated with rising prices, and vice versa, we might find prices significant in the consumption function due only to this distributional effect associated with a diminishing *GNP* gap.

Two points can be made to counter this hypothesis. First, balancing the increased income at the lower end of the income distribution is the well-known tendency for the profit share to rise in a cyclical upswing, shifting income to families with presumably lower-than-average consumption propensities.[19] Second, the correlation between price movements and the ratio of actual to potential real *GNP* is not all that clear in the period over which our consumption function was estimated.[20]

To test directly the hypothesis that our price terms only reflect cyclical effects, we reestimated the money-illusion consumption function (14) adding the natural logarithm of the ratio of actual real *GNP* to potential real *GNP* as a variable.[21] If the hypothesis is correct, the price term is merely a proxy for the closing of the *GNP* gap and inserting this new variable into the equation should greatly reduce the sum of the price coefficients and the significance of that sum, while assigning a significantly positive coefficient to the actual real *GNP*/potential real *GNP* variable.

Equation 3-2 of Table 3 gives the result of this test. The variable y/yp is the ratio of actual real *GNP* to potential real *GNP*. The rejection of the hypothesis that the price term is merely reflecting cyclical movements through a Phillips' curve mechanism is clear. While the cyclical *GNP* variable has a positive and nearly significant coefficient, the sum of the price coefficients is raised, not reduced, by the introduction of this cyclical variable.

Simultaneity Among Consumption, Income, and the CPI

Our consumption function is, of course, in reality part of a simultaneous system explaining aggregate consumption, income,

[19] See Kuh [12] for evidence on the cyclical behavior of the profit share.

[20] The period 1955 IV to 1958 II saw actual/potential *GNP* fall from 1.05 to 0.94 while the *CPI* rose from 93.5 to 100.0, a 7 percent increase. In the period 1961 I to 1965 IV, however, while actual/potential *GNP* rose from .94 to 1.03, the *CPI* rose from 103.9 to 111.1, still only 7 percent. See Kuh [13] for a recent criticism of the Phillips' curve explanation of price level determination.

[21] Potential *GNP* was computed following the Council of Economic Advisers' formulation. See [21, pp. 60–63].

TABLE 3—TESTS OF STATISTICAL PROBLEMS OF TREND, CYCLE, AND SIMULTANEITY

Equation	Dependent Variable	Constant	Independent Variables					Statistics		
			$\ln y/yt$		$\ln w/wt$	$\ln P/Pt$		R^2	$S.E.\times10^2$	$D.W.$
3-1	$\ln (c/ct)$	0.001 (0.001)	0.559 (0.055)		0.135 (0.040)	0.300 (0.110)		0.9496	.3337	1.430
			$\ln y$		$\ln w_t$	$\ln P$	$\ln y/yp$			
3-2	$\ln c$	−2.204 (0.191)	0.602 (0.056)		0.139 (0.036)	0.473 (0.049)	0.078 (0.048)	0.9984	.2900	1.887
			$\ln \hat{y}_t$	$\ln y_{t-i}$	$\ln w_t$	$\ln \hat{P}_t$	$\ln P_{t-k}$			
3-3	$\ln c$	−1.800 (0.137)	0.122 (0.102)	0.510 (0.102)	0.182 (0.041)	−0.204 (0.139)	0.566 (0.127)	0.9980	.3301	1.794
3-4	$\ln c$	−1.892 (0.131)	0.223 (0.092)	0.407 (0.089)	0.159 (0.041)		0.393 (0.042)	0.9977	.3434	1.751

and price determination. Because of the simultaneous relationships truly at work in this system, the error term in our consumption function (5) may be positively correlated with the contemporaneous values of y and P in that equation, biasing upward the estimates of the coefficients of these contemporaneous terms and downward the estimates of all the other coefficients.

We suspect an upward bias in the coefficient of $\ln y_t$ in the consumption function (14) because of the close connection of c_t and y_t through the usual national income accounts identity which appears in such simultaneous models. The coefficient of $\ln P_t$ is, however, less suspect because it is likely that prices in this quarter are mainly determined nonsimultaneously by events in previous quarters, through markup pricing procedures and the like. This leads us to consider $\ln P_t$ a predetermined, rather than a simultaneously determined, variable. Furthermore, there is no reason to expect past income, past prices, or current wealth to be determined simultaneously with current consumption.

To test the extent of such simultaneous equations bias, we employed the instrumental-variable technique.[22] Total labor income, ny, and the price level, P, were each regressed on a set of instruments, and then the resulting instrumental-variable estimates $\hat{y}_t=(\widehat{ny}/n)_t$ and $\hat{P}_t$ were used to replace y_t and P_t in estimating the basic equation (5).[23] First, we reestimated (5) replacing $\ln y_t, \ldots, \ln y_{t-6}$ by $\ln \hat{y}_t, \ln y_{t-1}, \ldots, \ln y_{t-6}$, and similarly replacing $\ln P_t, \ldots, \ln P_{t-6}$ by $\ln \hat{P}_t, \ln P_{t-1}, \ldots, \ln P_{t-6}$. The coefficients of income and price lagged t−1 to t−6 were estimated using the Almon technique with a third-degree polynomial lag distribution while $\ln \hat{y}_t$ and $\ln \hat{P}_t$ were entered separately into the regression. The resulting estimate is equation 3-3 in Table 3. The coefficients listed under $\ln y_{t-i}$ and $\ln P_{t-k}$ in 3-3 are the sums of the coefficients of the variables lagged 1 to 6 quarters.

In equation 3-3 the sum of the coef-

[22] See Malinvaud [14, pp. 604–08] for a discussion of the instrumental-variable technique.

[23] The instrumental-variable equations relating ny to current and lagged money supply, government expenditure, investment, and net exports, and relating P to current and lagged wage and profit rates are shown in the Appendix.

ficients of $ln\ y$ is 0.632, slightly less than the sum 0.661 in the consumption function estimate (14). The $ln\ \hat{y}_t$ coefficient alone is 0.122 as opposed to 0.274 in (14), although this comparison is not strictly legitimate since the $ln\ y_t$ coefficient in (14) is constrained by the Almon estimation procedure while the coefficient of $ln\ \hat{y}_t$ in 3-3 is not. The sum of the price coefficients in 3-3 is 0.362, again slightly lower than the sum of 0.418 of equation (14). The coefficient of $ln\ \hat{P}_t$ is insignificant and negative, while that of $ln\ P_t$ in (14) was insignificant and positive. The fit of 3-3 is worse than that of equation (14) although the Durbin-Watson statistic is slightly higher.[24]

Equation 3-4 shows the result of a second test of simultaneous equations bias. In this case, we tested only for income simultaneity, as equation (5) was reestimated replacing $ln\ y_t, \ldots, ln\ y_{t-6}$ by $ln\ \hat{y}_t, ln\ y_{t-1}, \ldots, ln\ y_{t-6}$, but including the original price lag $ln\ P_t, \ldots, ln\ P_{t-6}$. The resulting equation is similar to that of 3-3. The sum of the income coefficients is now 0.630, again insignificantly less than the 0.661 of the original (14), while the coefficient of $ln\ \hat{y}_t$ is 0.223, somewhat less than the coefficient 0.274 of $ln\ y_t$ in (14) but much higher than the coefficient of $ln\ y_t$ in 3-3. The sum of the price coefficients is insignificantly smaller in 3-4 than it was in (14).

These tests of simultaneous equations bias indicate that there may be a slight, but statistically insignificant, upward bias in our estimates of the coefficients of concurrent income and price in the money-illusion consumption function (14). Since the bias appears to be so small as to be insignificant it can probably be ignored quantitatively and certainly does not affect our principal qualitative conclusion: in the short run the price level has an independent effect on real consumption due to what is commonly called money illusion.

[24] Of course, the R^2 and $D.\ W.$ statistics and all the standard errors of 3-3 are equal to those of the true equation only asymptotically, since we are using y and P estimates which are only asymptotically equal to the true y and P values. But with R^2 values of .99+ on the instrumental variable equations, our standard errors are probably quite close to correct.

IV. *Concluding Comments*

The principal result of this paper—that a significant degree of money illusion appeared in the aggregate consumption function for the United States over the sample period—has a number of interesting implications for macroeconomic theory and policy. Two issues upon which it has an important bearing are the degree of stability of the economy and the nature of the aggregate labor supply function. These questions are discussed in Branson and Klevorick [6].

In closing, there are a number of lines of further investigation that our results suggest it might be fruitful to pursue. First, it would be interesting to disaggregate consumption expenditure and investigate such subaggregates as real personal consumer expenditures on durables and real personal consumer expenditures on nondurables and services using a money-illusion specification of the respective demand functions. Second, it would be most useful to introduce the money-illusion consumption function into a complete simultaneous equation model, observe its performance in such a model, and observe the implications for stability as viewed through simulation experiments. The results presented in this paper suggest that in constructing such macro-models, greater attention should be paid to the link between the price-wage sector and the expenditure sector.

APPENDIX

Instrumental-Variable Equations and Data Description

In Section III the money-illusion consumption function was reestimated with instrumental variable estimates of the *CPI*,

P_t, and per capita real net labor income, y_t, substituted for the actual series in the unlagged terms. While in principle the form of the instrumental equations used to construct the estimates should be irrelevant since the sole purpose of the technique is to break the simultaneity among c_t, y_t, and P_t (and clearly not to estimate a behavior function or structural equation for y_t or P_t), we will show here the estimated equations for $(ny)_t$ and P_t that were used to construct $(\widehat{ny})_t$ and $\hat{P}_t$. Then we will conclude with a description of the basic data series used in the study.

I. *Instrumental-Variable Equations for (ny) and P*

The instrumental variable equation for aggregate real net labor income, $(ny)_t$, has $(ny)_t$ as a function of current and lagged values of the money supply, real government expenditure, real gross private domestic investment, and real net exports. In linear estimating form the equation is

$$(ny)_t = \alpha + \sum_{i=0}^{I} \beta_i G_{t-i} + \sum_{j=0}^{J} \gamma_j I_{t-j} + \sum_{k=0}^{K} \delta_k X_{t-k} + \sum_{m=0}^{M} \eta_m M_{t-m} + \epsilon_t, \tag{A-1}$$

where G is government expenditure, I is gross private domestic investment, and X is net exports, all in billions of 1958 dollars from the *Survey of Current Business*, and M is the money supply, currency plus demand deposits in billions of dollars, from the *Federal Reserve Bulletin*.

As it turns out, the estimated version of (A-1) used to construct $(\widehat{ny})_t$, includes only current values of G and X, I lagged 0, 1, and 2 quarters, and M lagged 0–11 quarters with the coefficients estimated using a third degree Almon lag. The estimated equation is

$$\begin{aligned}(ny)_t = {} & -\underset{(12.30)}{238.53} + \underset{(0.109)}{0.423}G_t \\ & + \sum_{j=0}^{2} \gamma_j I_{t-j} + \underset{(0.160)}{0.838}X_t \\ & + \sum_{m=0}^{11} \eta_m M_{t-m}.\end{aligned} \tag{A-2}$$

$R^2 = .9976$; $S.E. = 1.78$; Mean = \$309.27 billion; $D.W. = 1.13$. Period of fit: 1955 I-1965 IV. The numbers in parentheses are standard errors.

The γ_i coefficients of lagged investment and the η_m coefficients of lagged money supply are shown in Table A-1. The coefficients of equation (A-2) were used to compute $(\widehat{ny})_t$.

The instrumental variable equation for P_t, the CPI, has P_t as a function of lagged wage rates, W, the average hourly earnings of manufacturing workers in dollars from the *Monthly Labor Review*, and the profit rate, R, the average rate of profit on stockholders' equity from the Federal Trade Commission *Quarterly Financial Reports on Manufacturing Corporations*. The estimated instrumental variable equation is

$$\begin{aligned}P_t = {} & \underset{(0.90)}{52.56} + \underset{(4.963)}{3.006}W_{t-1} + \underset{(4.271)}{9.860}W_{t-2} \\ & + \underset{(2.455)}{7.973}W_{t-3} + \underset{(3.711)}{2.851}W_{t-4} \\ & - \underset{(0.046)}{0.138}R_{t-1}\end{aligned} \tag{A-3}$$

$R^2 = .9929$; $S.E. = 0.47$; Mean $= 101.95$; $D.W. = 0.83$. Period of fit: 1955 I-1965 IV. The numbers in parentheses are standard errors.

The coefficients of equation (A-3) were used to compute the $\hat{P}_t$ series used in the text.

II. *Data Description*

The data used for real consumption per capita, c, are essentially aggregate real consumption expenditures on nondurables and services plus depreciation and imputed interest on durables, divided by population, n. The imputed interest on durables represents consumers' use of durables' services. Real net labor income per capita, y, is aggregate employees' compensation plus an imputed proportion of proprietors' income plus trans-

TABLE A-1—COEFFICIENTS OF I_{t-j} AND M_{t-m} IN INSTRUMENTAL VARIABLE EQUATION (A-2) FOR ny

Lag (j, m)	Coefficient of	
	I_{t-j}	M_{t-m}
0	0.423 (0.092)*	0.582 (0.145)
1	0.155 (0.110)	0.290 (0.067)
2	0.204 (0.094)	0.123 (0.056)
3	0 ·	0.054 (0.068)
4	· ·	0.062 (0.068)
5	· ·	0.121 (0.056)
6	· ·	0.209 (0.041)
7	· ·	0.302 (0.038)
8	· ·	0.375 (0.052)
9	· ·	0.405 (0.065)
10	· ·	0.368 (0.066)
11	· ·	0.241 (0.047)
12	0	0

* The numbers in parentheses are the standard errors of the coefficients.

fer receipts less employees' social insurance contributions and state, local, and federal tax liabilities on labor income, deflated by the consumption deflator and divided by n. Real wealth per capita, w_t is the aggregate net worth of households, including liquid assets, consumer durables, and housing, deflated by the consumption deflator and divided by n. These three series of wealth components, all in billions of 1958 dollars, quarterly at annual rates, are updated versions of the annual series used by Ando and Modigliani [2]. Population, n, is total *U.S.* population in millions. The consumption and income data and the consumption deflator were provided to us by Harold Shapiro, and the wealth and population data were provided by Albert Ando. The *CPI* data are quarterly averages of the monthly figures published in the *Survey of Current Business*.

The authors will be happy to make the entire set of data available upon written request.

REFERENCES

1. S. ALMON, "The Distributed Lag Between Capital Appropriation and Expenditures," *Econometrica*, Jan. 1965, *33*, 178–96.
2. A. ANDO AND F. MODIGLIANI, "The 'Life-Cycle' Hypothesis of Saving: Aggregate Implications and Tests," *Amer. Econ. Rev.*, Mar. 1963, *53*, 55–84.
3. J. J. ARENA, "Capital Gains and the 'Life-Cycle' Hypothesis of Saving," *Amer. Econ. Rev.*, Mar. 1964, *54*, 107–11.
4. ———, "The Wealth Effect and Consumption: A Statistical Inquiry," *Yale Econ. Essays*, fall 1963, *3*, 251–303.
5. C. W. BISCHOFF, "Elasticities of Substitution, Capital Malleability, and Distributed Lag Investment Functions," presented at Econometric Society, December 28, 1966.
6. W. H. BRANSON AND A. K. KLEVORICK, "Money Illusion in Static and Dynamic Macroeconomic Models," unpublished manuscript, Sept. 1968.
7. ——— AND L. C. THUROW, "Final Report on Demography and Demand," unpublished Council of Economic Advisers memorandum, July 26, 1965.
8. R. FERBER, *A Study of Aggregate Consumption Functions*. New York 1953.
9. Z. GRILICHES, "Distributed Lags: A Survey," *Econometrica*, Jan. 1967, *35*, 16–49.
10. E. J. KANE AND A. K. KLEVORICK, "Absence of Money Illusion: A *Sine Qua Non* for Neutral Money?" *J. Finance*, Sept. 1967, *23*, 419–23.

11. L. KOYCK, *Distributed Lags and Investment Analysis*. Amsterdam 1952.
12. E. KUH, "Income Distribution and Employment Over the Business Cycle," in J. S. Duesenberry, G. Fromm, L. R. Klein, and E. Kuh, eds., *The Brookings Quarterly Econometric Model of the United States*. Chicago 1965, pp. 227–78.
13. ———, "A Productivity Theory of Wage Levels—An Alternative to the Phillips Curve," *Rev. Econ. Stud.*, Oct. 1967, *34*, 333–60.
14. E. MALINVAUD, *Statistical Methods of Econometrics*. Chicago 1966.
15. F. MODIGLIANI AND R. BRUMBERG, "Utility Analysis and Aggregate Consumption Functions: An Attempt at Integration," unpublished manuscript.
16. ———, "Utility Analysis and the Consumption Function: An Interpretation of Cross-section Data," in K. K. Kurihara, ed., *Post Keynesian Economics*. New Brunswick 1954, pp. 388–436.
17. ——— AND R. SUTCH, "Innovations in Interest Rate Policy," *Amer. Econ. Rev.*, Proc., May 1966, *56* 178–97.
18. D. PATINKIN, *Money, Interest, and Prices*, 2nd ed., New York 1965.
19. J. H. POWER, "Price Expectations, Money Illusion, and the Real-Balance Effect," *J. Polit. Econ.*, Apr. 1959, *67*, 131–43.
20. P. A. SAMUELSON, *Foundations of Economic Analysis*. Cambridge, Mass., 1947.
21. COUNCIL OF ECONOMIC ADVISERS, *The Annual Report of the Council of Economic Advisers*, Washington, Feb. 1968
22. National Bureau of Economic Research, *The Price Statistics of the Federal Government; Review, Appraisal, and Recommendations*. New York 1961.

A Theory and Test of Credit Rationing

By Dwight M. Jaffee and Franco Modigliani*

Nonprice credit rationing by commerical banks and other intermediaries has attracted a good deal of attention in recent years, in part because of the role assigned to this phenomenon by the Availability Doctrine as developed by Robert Roosa [9] and others in the years immediately following World War II[1]. It is by now generally agreed that credit rationing, if it could be shown to be empirically widespread, would have important implications for an assessment of the effectiveness and timeliness of monetary policy as well as for our understanding of its *modus operandi*. But significant disagreement still exists whether credit rationing is consistent with rational bank behavior and whether it is an important empirical phenomenon. These two issues essentially define the goal of this study; namely, to provide affirmative answers to the following questions:

(1) Is it rational for commercial banks to ration credit by means other than price?
(2) Can credit rationing be measured? If so, are there significant variations in rationing over time and can these variations be accounted for?

With respect to the second question, empirical research has been hampered by the almost insolvable problem of directly measuring credit rationing. This has led to the use of proxy variables in the form of indicators of tight money such as interest rate levels or changes in interest rates. It is difficult to obtain conclusive results with such variables, however, since one cannot then really differentiate credit rationing from other symptoms of tight money. In this study, by contrast, we are able to derive and exhibit an operational proxy for credit rationing based explicitly on a theory of rational lender behavior.

The major contributions concerned with the first question have been provided in a series of complementary studies by Donald Hodgman [3], Merton Miller [8], and Marshall Freimer and Myron Gordon [1].[2] Although criticism of a technical nature has been raised with respect to the first two studies, the most recent of these works by Freimer and Gordon, provides a complete statement of the model. Consequently, it may seem surprising that even the rationality of credit rationing is still considered a debatable issue. The source of this paradox, we believe, is essentially that the authors have not addressed themselves to the relevant question. Before elaborating on this point, however, it is important to precisely define credit rationing.

In line with the generally accepted terminology, we propose to define credit ra-

* The authors are, respectively, assistant professor of economics at Princeton University and professor of economics and industrial management at Massachusetts Institute of Technology. The study was carried out in conjunction with the Federal Reserve–Massachusetts Institute of Technology econometric model sponsored by the Social Science Research Council with support from the Board of Governors of the Federal Reserve System. The authors also wish to express their thanks to other participants in the MIT-FRB project, and to Franklin M. Fisher, Stephen M. Goldfeld, Burton G. Malkiel, and Gerald A. Pogue for helpful criticism and suggestions. The views expressed in the paper are those of the authors alone.

[1] Credit rationing is discussed in this study only within the institutional framework of the commercial banking industry. Nonetheless, large sections of the paper, and particularly the theoretical model, could be applied equally well to other financial intermediaries.

[2] This list is not inclusive and references to other works can be found in the studies cited. A second aspect of the mechanism of credit rationing based on "customer relationships" is developed by Hodgman [3] and Kane and Malkiel [6] and is discussed further below.

tioning as a situation in which the demand for commercial loans exceeds the supply of these loans at the commercial loan rate quoted by the banks. Thus credit rationing is an excess demand for commercial loans at the ruling commercial loan rate. In addition, it is helpful to distinguish two forms of credit rationing depending on the status of the commercial loan interest rate. *Equilibrium rationing* is defined as credit rationing which occurs when the loan rate is set at its long-run equilibrium level. *Dynamic rationing* is defined as credit rationing which may occur in the short run when the loan rate has not been fully adjusted to the long-run optimal level.

This definition serves to bring into focus the basic challenge which must be met in order to establish the rationality of equilibrium rationing, namely: can it ever be rational for the bank to limit the loan to less than the amount demanded by the borrower when both the loan rate and the loan granted are chosen optimally, that is, to maximize profits. The problem of dynamic rationing differs only in that the commercial loan rate is set at levels consistent with short-run profit maximization.[3] In either case, three elements enter into the problem: the demand for loans, the supply of loans, and the determinants of the commercial loan rate. The shortcoming of the earlier studies already cited is that they concentrate on the determinants of the quantity supplied by lenders while neglecting the other two elements.[4]

[3] Previous discussions of credit rationing have studied the two forms of rationing independently: the theoretical justification of credit rationing was directed only at the equilibrium form; the empirical significance was considered only with respect to dynamic rationing. In our analysis we endeavor to integrate the two because our theory and the related empirical tests stress their common origin.

[4] The papers by Hodgman [3] and Miller [8] and the discussion of "weak credit rationing" in Freimer and Gordon [1] essentially just omit any reference to demand and the determinants of the rate. The discussion of "strong credit rationing" by Freimer and Gordon states the question properly but then assumes the answer, as shown in fn. 12.

It should be clear that information on the supply curve alone will generally not be sufficient to derive implications about credit rationing. For this reason, the development of our theoretical model of credit rationing integrates the demand for loans and the determinants of the loan rate with the supply of loans. The development of this model is given in Sections I and II of the paper. In Section I, analytic propositions concerning rationing are derived under alternative assumptions about market competition. In Section II, these results are then interpreted in the light of the institutional structure of competition in the commercial banking industry of the United States to provide the complete theory. Section III describes how our theory of credit rationing can be used to derive an operational credit rationing proxy. Section IV contains empirical tests of the theory as set out in Sections I and II using the credit rationing proxy derived in Section III.

I. *Some Analytic Propositions*

1. *The Bank's Optimal Loan Offer Curve*

The first set of propositions to be developed are concerned only with the bank's supply curve or offer curve for commercial loans. This offer curve is derived by generalizing the results obtained by Freimer and Gordon [1] for a rectangular density function of possible returns to the bank on commercial loan contracts. We shall then proceed to include a demand function and derive the implications of alternative forms of competition.

To start, consider a banker facing a large number of customers each wishing to finance its investment projects. As in the earlier literature, we define the outcome of the customer's projects as the firm's end of period value, denoted by x. We further assume that the bank views x as a random variable and summarize in the density function $f_i[x]$ the bank's subjective evalua-

tion of the probability of different outcomes. In general, the density function $f_i[x]$ will be affected by the size of the customer's investment which, in turn, may be expected to depend on the size of the loan granted. For expository convenience, our formal analysis here will proceed under the restrictive assumption that the size of the project is fixed and therefore $f_i[x]$ can be taken as independent of the loan size granted. This assumption implies that the firm has alternative means of finance which can be used to complement the bank credit. It can be shown, however, that all the major conclusions reached here can be generalized to the case where the size of the project is not independent of the loan granted, provided the investment opportunity of the customer is subject to decreasing (expected) returns.[5]

The expected profit of the bank from the ith customer loan, P_i, is a function of the size of the loan made L_i, the loan rate r_i, and the density function f_i. It is helpful to assume the existence of a sure minimum outcome k_i and a maximum possible outcome K_i for the projects such that:

$$f_i[x] = 0 \qquad \text{for } x < k_i \quad \text{or} \quad x > K_i$$

Now let $R_i = 1 + r_i$ be the interest rate factor and thus R_iL_i will be the total amount of the contract repayment (interest plus principal). Then the contribution of L_i to the bank's expected profits can be written as the difference between the total expected repayment by the firm and the total opportunity cost, i.e.:

$$(1) \qquad P_i = P_i[R_iL_i] = R_iL_i \int_{R_iL_i}^{K_i} f_i[x]dx + \int_{k_i}^{R_iL_i} xf_i[x]dx - IL_i$$

The first term in this expression represents the gross receipts of the bank if the outcome x is sufficiently favorable to enable the firm to repay the agreed amount R_iL_i in full. The second term denotes the receipts if the outcome of the project falls short of the contracted amount. In this case, we assume the bank receives the entire outcome x, whatever it might be.[6] The last term represents the bank's opportunity cost, where $I = 1 + j$ and j is the opportunity rate. The rate is for the moment assumed constant and independent of the loan contract on the premise that the bank has unlimited access to a perfect capital market.[7]

The expected profit function (1) can be further simplified by adding and subtracting

$$R_iL_i \int_{k_i}^{R_iL_i} f_i[x]dx$$

and then integrating the second term of that expression by parts. This yields:

$$(2) \qquad P_i = P_i[R_iL_i] = (R_i - I)L_i - \int_{k_i}^{R_iL_i} F_i[x]dx$$

where

$$F_i[A] = \int_{k_i}^{A} f_i[x]dx$$

[5] Freimer and Gordon [1] conclude that there is a significant difference between the fixed size investment and variable size investment assumptions. However, this conclusion is based on an inconsistent definition of opportunity cost. In Jaffee [5] the essential equivalence of these two assumptions, using a proper definition of opportunity cost, is derived. See also fn. 11.

[6] In such cases of at least partial default, it is reasonable that the bank would incur some collections which should be deducted from the outcome x. While this refinement is neglected here, it can be shown that the character of our conclusions would not change if it were taken into account. See Jaffee [5] and Miller [8]. A similar remark applies to the cost of making and servicing a loan.

[7] The existence of well-developed markets in Federal Funds and Certificates of Deposits (*CDs*) make this assumption reasonable in normal periods. If these markets cease to operate in tight money periods, for example when the Regulation Q ceiling hinders the issue of new *CDs*, then the shadow price of funds would have to replace the market indicator as the opportunity cost. This may lead to dynamic effects, and these are considered below in the discussion of dynamic rationing (Section II.3) and in the empirical tests.

is the probability that x will be less than a.

An optimal loan to a customer is defined as the loan size which maximizes the bank's expected profits from that customer for a given loan interest rate.[8] The bank's offer curve for that customer is then the set of optimal loans corresponding to alternative possible loan rates. This offer curve can be derived from the first-order condition for the maximization of expected profit:

$$\text{(3)} \quad \frac{\partial P_i[R_iL_i]}{\partial L_i} = R_i(1 - F[R_iL_i]) - I = 0$$

This condition can be usefully rewritten in the form:

$$F_i[R_iL_i] = 1 - \frac{I}{R_i} = \frac{r_i - j}{1 + r_i}$$

Since the quantity $F_i[R_iL_i]$ is precisely the probability of default, (3) admits the following simple interpretation: the optimal loan is such that the probability of default is equal to the excess of the loan rate over the opportunity cost, normalized by the loan rate factor $R_i = 1 + r_i$.[9]

The offer curve can now be defined as the implicit solution to (3) for L_i in terms of R_i subject to the nonnegativity condition $L_i \geq 0$. We will denote this solution by $\hat{L}_i = \hat{L}_i[R_i]$. From (3) we can deduce several properties of this offer curve which will be used in developing our argument.

[8] The formulation in terms of expected profits assumes, of course, a linear utility function for the bank. This makes it unnecessary to take into account higher order moments of the distribution of outcomes for the firm's projects and covariance of profits between customers. The rationale for this assumption is that banks service a large number of relatively diverse customers. In addition, a utility function with explicit risk aversion would leave unclear whether the existence of credit rationing in the model arises from the structure of the model or simply from the utility function.

[9] Since the empirically observed spread between the loan rate and the opportunity rate at which banks can secure or invest funds is typically small, this condition implies that banks should tend to assume quite modest default risk. This conclusion seems to be in agreement with bank loss experience on commercial loans.

PROPOSITION 1. The optimal loan offer curve defined by condition (3) (and drawn in Figure 1 has the following properties:

(1.1) $\hat{L}_i = 0 \qquad$ for $R_i < I$

(1.2) $0 \leq \hat{L}_i \leq k_i/I \qquad$ for $R_i = I$

(1.3) $R_i\hat{L}_i \leq K_i \qquad$ for all R_i

(1.4) $\lim_{R_i \to \infty} \hat{L}_i = 0$

Proposition (1.1) follows from the nonnegativity condition and condition (3). Since the marginal expected profit of an additional loan $(\partial P_i/\partial L_i)$ is negative for all positive loan sizes whenever $R_i < I$, the best the bank can do under this condition is to extend no loan at all. Similarly, when $R_i = I$, condition (3) can hold only if $F_i[R_iL_i] = 0$; this implies $R_iL_i = IL_i \leq k_i$ which is equivalent to proposition (1.2). This means that the offer curve is a vertical line segment when $R_i = I$. In fact, in the case with no uncertainty in which F_i is identically zero, the offer curve is nothing more than this vertical line.

The logic of proposition (1.3) is that for any given interest rate factor R_i, the bank will not receive additional income from extending loans in amounts which exceed the solution of $R_iL_i = K_i$, because K_i is the maximum amount the firm could conceivably earn. Furthermore, loans cost the bank the opportunity rate I, and hence all solutions must satisfy the condition of proposition (1.3). *Thus the optimal loan is finite no matter how high the interest rate offered.*[10] From proposition (1.3) follows immediately proposition (1.4). It implies that as the loan rate grows larger and larger, the optimal loan does not follow course; to the contrary, at least after some point, the optimal loan will begin to decline as the rate grows and will eventually

[10] Although the maximum is finite, there may be multiple local maxima for the loan offer curve. In the discussion which follows, we abstract from this complication since it does not affect the results.

approach zero as the rate grows beyond bounds. The common sense of this surprising implication can be understood from the following considerations: (*i*) by making the size of the contracted repayment R_iL_i sufficiently large, default becomes virtually certain, and hence the bank can count on becoming the owner of all the net activities of the firm; (ii) as R_i rises, the amount L_i that the bank needs to invest to achieve this result grows smaller and smaller and approaches zero as R_i tends to infinity.[11]

Two other properties of the offer curve are also worth developing.

PROPOSITION 2. For a given interest factor R_i, expected profits decrease monotonically as the loan size varies from the optimal size in either direction.

The proof follows directly from (3), since

$$\frac{\partial^2 P_i[R_iL_i]}{\partial L_i^2} = -R_i^2 f_i[R_iL_i] < 0 \tag{4}$$

$$\text{for } k_i < R_iL_i < K_i$$

This also shows that the solution given by (3) is, indeed, a global maximum.

PROPOSITION 3. Expected profits increase along the offer curve for successively higher interest rate factors.

To derive this result, first substitute (3) into (1), which allows us to write the profit function along the offer curve as:

$$P_i[R_iL_i] = \int_{k_i}^{R_i\hat{L}_i} xf[x]dx \tag{5}$$

But then,

$$\frac{dP_i[R_i\hat{L}_i]}{dR_i} = R_i\hat{L}_i f_i[R_i\hat{L}_i]\left(\hat{L}_i + R_i\frac{d\hat{L}_i}{dR_i}\right)$$

and by implicitly differentiating (3)

$$\hat{L}_i + R_i\frac{d\hat{L}_i}{dR_i} = \frac{1 - F_i[R_i\hat{L}_i]}{R_if_i[R_i\hat{L}_i]}$$

and thus

$$\frac{dP_i[R_i\hat{L}_i]}{dR_i} = \hat{L}_i(1 - F_i[R_i\hat{L}_i]) > 0 \tag{6}$$

$$\text{for } k_i < R_i\hat{L}_i < K_i$$

which proves the proposition. An obvious and reasonable implication of this proposition is that the bank will obtain its maximum potential profits when allowed to charge an infinite interest rate. Of course, this only serves to emphasize that the offer curve is defined independently of the demand for loans and thus many points on the offer locus may not prove feasible for the lender.

2. *The Banker as a Discriminating Monopolist*

The optimal loan offer curve just derived can be interpreted as the bank's supply curve for the bank customer. It is interesting that this supply curve is "backward bending" as shown in Figure 1. In fact, Freimer and Gordon [1] base part of their argument for credit rationing on this characteristic of the supply curve alone.[12] Our

[11] A similar result holds for the variable size investment case. Although the optimal loan offer remains positive as the interest rate approaches infinity, the loan size approaches a finite asymptote. In contrast, Freimer and Gordon [1, p. 407] conclude that the optimal loan approaches infinity as the interest rate goes to infinity, because they only consider projects with an expected value exceeding the opportunity cost; but this contradicts the meaning of opportunity cost; see fn. 5.

[12] Reference is to the Freimer and Gordon [1] case of weak credit rationing. The terminology is misleading because one would normally not equate a backward bending supply curve with rationing. Freimer and Gordon also consider what they call "strong credit rationing," which conceptually is closely akin to our equilibrium rationing. For this case they argue that a bank sets a conventional interest rate (6 percent to be exact) and then grants loans to all customers up to the amount indicated by their respective bank offer curves. This analysis is inadequate for two reasons. First, the optimality of a 6 percent rate is essentially just assumed; banks do not charge rates below 6 percent because of convention; banks do not charge rates above 6 percent because their numerical examples suggest

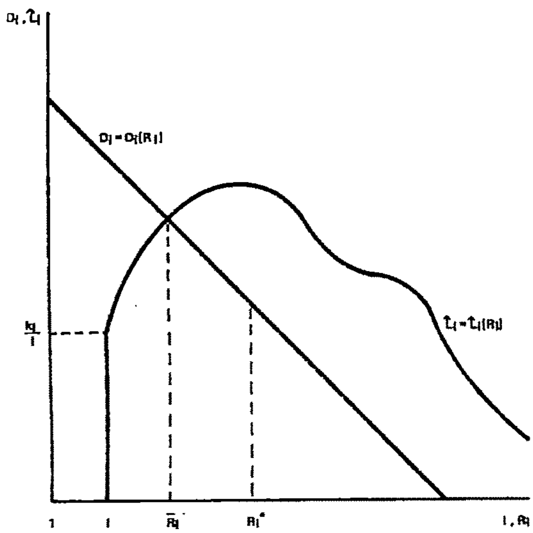


FIGURE 1

definition of credit rationing, however, obviously necessitates taking into account the demand for loans by the bank customer. Such a demand curve for the ith bank customer, denoted hereafter by $D_i = D_i[R_i]$, has been drawn in Figure 1. We assume a downward sloping demand curve, which vanishes for a sufficiently high interest rate factor and is finite even for a zero interest rate (or interest factor equal to unity). The boundary conditions for the demand curve are based on the characteristics of the underlying investment projects. At a zero interest rate the firm's demand for loans is bounded by the size of the investment projects, while the existence of some alternative means of financing suggests that at some sufficiently high interest rate the demand will become zero. The negative slope of the demand curve can be derived from the assumption that the firm has only limited access to these alternative means of finance. The implied limit on competition between banks for the customer's business is discussed further in Section II.

The conditions under which rational credit rationing will occur can now be seen with the aid of Figure 1. The ith firm's demand curve and the bank's offer curve to that customer are assumed to intersect at the interest rate factor $\bar{R}_i$.[13] If the bank chooses to charge an interest factor greater than $\bar{R}_i$, rationing will not occur since the loan demand is less than the loan offer at such an interest factor. In fact, if the firm were willing to accept a larger loan (which it is not by definition of the demand curve), the bank would increase its expected profits by providing such a loan (see proposition 2). On the other hand, if the bank chooses to charge an interest factor less than $\bar{R}_i$, then rationing will occur since in this region the bank's optimal loan offer is less than the amount demanded by the customer. The bank would only reduce its expected profits by increasing the loan offer to meet the demand. Thus the question of the rationality of credit rationing can be reduced to a consideration of the optimal rate factor to be charged by the bank, and its relation to $\bar{R}_i$.

Two critical variables enter into the bank's selection of the loan rate. First there is the question of the time horizon: in the long run, a rational banker would select the rate which maximizes his expected profits; but other constraints may preclude immediate full adjustment in the short run. This leads, of course, to our distinction between the cases of equilibrium rationing and dynamic rationing as already defined. We start by considering only the equilibrium case since it has been

that *customer* utility maximization occurs at lower rates. Secondly, even if the 6 percent rate is accepted as optimal, we have no way of knowing from their analysis whether rationing actually occurs, because the demand curve is never shown.

[13] The existence of at least one point of intersection is assured by the boundary conditions on the demand curve and proposition (1.4). Throughout the remainder of the paper we shall assume that, in fact, only one such point exists. Equivalent results can be obtained for the case of multiple intersections, but only at the cost of more complexity in the analysis.

the center of the theoretical discussion and since the short-run dynamic case is easily derived from the equilibrium case.

The second important variable influencing the bank's choice of rate is the nature of market competition. Because the degree of competition turns out to be a critical factor in determining the existence of rationing, it is worthwhile considering several different regimes from a purely analytic standpoint.

The first regime considered is the simple case of a discriminating monopolist. It is assumed that the bank maximizes its expected profits with respect to each customer separately and is free to charge each customer a different interest rate. Thus, we can take the bank's solution for the ith customer as typical:

PROPOSITION 4. Let R_i^* be the rate factor which maximizes the bank's expected profits when the bank is acting as a discriminating monopolist. Then:

(4.1) $R_i^* \geq \overline{R}_i$, which implies

(4.2) Credit rationing is not profitable for a banker acting as a discriminating monoplist.

The proof of this proposition is easily derived from Figure 1 and the properties of the offer curve. The bank will always charge an interest factor at least as high as $\overline{R}_i$ since; (i) expected profits increase along the offer curve for successively higher interest rates (proposition 3) and (ii) the loan $L_i(\overline{R}_i)$ is feasible by the definition of $\overline{R}_i$. We next observe that, in view of proposition 2, for any rate equal to, or larger than $\overline{R}_i$, the bank's profit will be higher the closer the loan is to the corresponding point on the offer curve. But since the demand curve places a ceiling on the feasible loan size, the bank will find the optimal rate R_i^* by maximizing profit along the demand curve constraint, which means that credit rationing will not be profitable. Indeed, as in the standard theory of monopoly under certainty, at the rate R_i^* the bank would be glad to lend more than the customer is prepared to take.

3. *The Banker Must Charge All Customers a Uniform Rate of Interest*

Now suppose that the banker is constrained to charge all customers the same rate though he can choose that rate freely and can also decide on the size of the loan to be granted each customer. We will show that under these conditions credit rationing may (and very frequently will) be profitable, i.e. at the common optimal interest rate, for some customers the most profitable loan for the bank to supply is less than the amount demanded.

To establish this proposition it is convenient to deal first with a subsidiary problem. Suppose that the bank faces only two customers and, for the moment, rule out credit rationing by requiring that the bank *must* satisfy the customers' demand at the chosen common rate.

The bank's expected profits under these conditions can be written as:

$$P = P_1[RD_1] + P_2[RD_2] \tag{7}$$

where R is the common rate factor charged both customers and the constraint of satisfying both demand functions is implicit in the notation. By differentiating this profit function with respect to the interest factor, we can obtain the first-order condition for the optimal interest factor, say R^*.

An important property of R^* can be obtained, however, without explicit reference to this first order condition. Let R_1^* and R_2^* be the optimal rate factors a discriminating monopolist banker would charge customers 1 and 2, respectively, and assume, without loss of generality, that $R_1^* < R_2^*$. It follows from this definition that the expected profit from each cus-

tomer must be a concave function of R in the neighborhood of R_1^* and R_2^*, respectively. We shall go somewhat further and assume that the concavity of each expected profit function holds for all $R > 1$.[14] We can then establish:

PROPOSITION 5. If R^* is the common rate factor that maximizes the bank's expected profit, we must have:

$$R_1^* \leq R^* \leq R_2^*$$

This result can be verified through use of a proof by contradiction. The assumption of concave expected profit functions implies that expected profits decrease monotonically as the absolute value of the spread between the actual rate to a customer and the discrimination monopolist rate increases. Thus, if the bank chose a common rate factor R^* that was less than both R_1^* and R_2^*, it would find that expected profits could be increased by increasing the rate factor at least to the level of R_1^*, thus contradicting the assumption that the original rate factor was optimal. Essentially the same argument shows that $R^* > R_2^*$ also leads to a contradiction.

We can now relax the restriction that the bank must satisfy both customers' demand function and show that, under these conditions, it will never pay to ration customer 1 but it may very well pay to ration customer 2. Specifically, we can establish:

PROPOSITION 6. For the common rate regime

(6.1) $R^* \geq R_1^* \geq \overline{R}_1$ implying it is not profitable to ration customer 1.

(6.2) $R_2^* \geq R^* \gtrless \overline{R}_2$, implying that $R^* < \overline{R}_2$ is possible, in which case, credit rationing is profitable.

Proposition (6.1) follows directly from propositions 5 and 4. As for establishing (6.2), we need only exhibit a concrete example in which the condition holds. That it is, in fact, quite easy to construct such examples can be seen from the following considerations. First, going back to proposition 5, one can readily establish that the position of R^* within the range R_1^* to R_2^* depends on the relative size of the two customers (as measured, say, by the size of the loan demanded for any R in the critical range), and on the elasticity of the two demand curves. In particular, R^* can be made arbitrarily close to R_1^* by assuming that customer 1 is sufficiently larger than customer 2, and/or by assuming that the demand curve for customer 1 is sufficiently inelastic in the range of rates above R_1^*. By the same token, $\overline{R}_2$ can be made arbitrarily close to R_2^* by assuming a sufficiently elastic demand curve for customer 2. Thus, by appropriate choice of these functions, one can readily construct situations where R^* is lower than $\overline{R}_2$. The above construction also provides an interesting interpretation of proposition 6. The constraint of charging both customers the same rate R^* forces the bank to charge customer 1 a rate which is too high relative to R_1^* and hence, the customer is not rationed. On the other hand, the bank is forced to charge customer 2 a rate which is too low relative to R_2^*. If the rate is sufficiently low relative to $\overline{R}_2$, that is $R^* < \overline{R}_2$, then the second customer will be rationed.

The possibility of credit rationing also adds another dimension to the problem. In the case in which rationing does not occur, R^* is the optimal common rate charged by the bank and the customers receive loans

[14] The assumption of global concavity is not a necessary condition for proposition 6, to be proven below; the proof can be carried out with weaker assumptions. We have chosen to assume global concavity, however, because it leads, via proposition 5, to a particularly interesting demonstration of proposition 6.

of D_1 [R^*] and D_2 [R^*], respectively. But when customer 2 can be profitably rationed at R^*, the very existence of rationing changes the conditions of the problem and R^* need no longer be the optimal rate. Instead, there will exist a more general optimum optimorum rate, say $\hat{R}$, which yields the maximum expected profits after allowing for credit rationing of customer 2. $\hat{R}$ will just equal R^* in cases in which rationing is not profitable, but will generally differ from R^* when rationing is profitable. The fact that $\hat{R}$ may differ from R^* creates some difficulties in deriving the comparative static properties of the model, as will be apparent below. However, it does not affect our basic conclusion; proposition 6 remains valid even if R^* is replaced with the true optimum $\hat{R}$. For, as one can readily verify, $R^* \leq \overline{R}_2$ implies $\hat{R} \leq \overline{R}_2$, and conversely.[15]

Before proceeding to the generalizations of the propositions developed so far, it is worthwhile considering the special case of a risk free firm, that is a customer for whom the bank's subjective evaluation of default risk is zero.

PROPOSITION 7. Neither a banker acting as a discriminating monopolist, nor a banker charging all customers the same rate, will ration a risk free customer.

The result of proposition 7 for the case of a discriminating monopolist is simply a special case of proposition 4 and follows directly from that proposition. To prove the proposition for a banker charging a common rate factor, it is important to recall from proposition 1 that the bank's offer curve for a risk free customer is a vertical line at $R=I$. Furthermore, it must be true that R^* (or $\hat{R}$) is greater than I if the bank's expected profits are to be positive. But this implies that $R^* \geq \overline{R}$ for the risk free customer, which on the basis of proposition 6 rules out the possibility of credit rationing.

We can now proceed to generalizations of the propositions developed in this part. First, consider a banker facing n customers and constrained to charge all n customers the same loan rate. Again, let R_i^* ($i=1,2,\ldots,n$) denote the rate the bank would charge if acting as a discriminating monopolist and let $\hat{R}$ be the common "optimum optimorum" rate the bank charges when allowing credit rationing. Let us number customers in ascending order with respect to the monopolist rate; that is $R_i^* \geq R_{i-1}^*$ for $i=2,3,\ldots,n$. Proposition 5 can then be generalized to:

PROPOSITION 5′. The optimal common rate $\hat{R}$ charged all customers must lie between the rate charged customer 1 and the rate charged customer n when the bank is acting as a discriminating monopolist. Formally, there exists an integer j, $2 \leq j \leq n$, such that $R_j^* \geq \hat{R} \geq R_{j-1}^*$.

Similarly, proposition 6 can be generalized to:

PROPOSITION 6′.

(6.1′) For any customers i such that $\hat{R} \geq R_i^*$, we have $\hat{R} \geq \overline{R}_i$ implying that rationing of these customers is not profitable.

(6.2′) In the case of customers for whom $\hat{R} < R_i^*$, we have $\hat{R} \gtrless \overline{R}_i$ implying that rationing of some of these customers may be profitable.

The interpretation and proof of these

[15] We have previously established that $R^* \geq \overline{R}_2$ implies that R^* is already the optimum optimorum or $R^* = \hat{R}$; hence $R^* \geq \overline{R}_2 \rightarrow \hat{R} \geq \overline{R}_2$. Similarly $\hat{R} \geq \overline{R}_2$ implies that even when rationing is allowed, it is not profitable and therefore $\hat{R}$ must coincide with R^*, the maximum subject to the constraint that rationing is not permissible; that is $\hat{R} \geq \overline{R}_2 \rightarrow \hat{R} = R^*$ and therefore $\hat{R} \geq \overline{R}_2 \rightarrow R^* \geq \overline{R}_2$. But these two propositions together imply $R^* \geq \overline{R}_2 \Leftrightarrow \hat{R} \geq \overline{R}_2$, which in turn implies $R^* \leq \overline{R}_2 \Leftrightarrow \hat{R} \geq \overline{R}_2$.

propositions follows directly from the corresponding propositions for the case of two customers, and accordingly are not repeated here.

4. *Generalization to m Separate Customer Classes*

It is now easy to extend our conclusions about the rationality of credit rationing to the case in which the bank can assign customers to any one of, say, m classes, where within each class the bank must charge a single uniform rate. The principles which govern the assignment of customers to classes and the choice of the rate for each class can be readily inferred from the previous analysis:

(i) If the profit functions are all concave in the relevant range, then the optimal classification will be achieved by dividing the entire range of the set of R_i^* (the rate charged customer i when the bank acts as a discriminating monopolist) into m intervals and assigning to the same class all customers whose R_i^* falls in a given interval. If the profit functions are not concave, the principle for optimal classification becomes more complex; but in any event, it is clear there will exist a set of optimal group rates and a corresponding optimal classification for the customers, and that furthermore, each class will contain customers with different R_i^*s, as long as the number of customers exceeds the number of classes.

(ii) The optimum rate for any given class j, say $\hat{R}_j$, must fall somewhere between the smallest and the largest R_i^* of the customers in that class.

(iii) It will not be profitable to ration customers whose R^* is smaller than the group rate $\hat{R}_j$, but it may pay to ration those for whom R_i^* exceeds $\hat{R}_j$. In particular, rationing will occur whenever $\bar{R}_i > \hat{R}_j$.

(iv) The likelihood that it will be profitable to ration at least some customers in a class will be positively related to the heterogeneity of the R_i^* of the customers in that class and hence the likelihood of rationing will be inversely related to the size of m. Indeed, if m is allowed to be as large as the number of customers, that is, n, then the bank will be in the position of a discriminating monopolist and credit rationing will not occur.

There remains now to draw the implications of these results by combining the propositions developed so far with a number of considerations arising from the nature of competition in the banking industry.

II. *Competition in Banking and Credit Rationing*

1. *The Nature of Competition in Commercial Banking and its Implications for Credit Rationing.*

We have shown in Section I that a single bank, free to discriminate between borrowers by charging each customer its monopolist rate R_i^*, would not ration credit. A similar conclusion holds even if there are many banks, as long as they act collusively to maximize joint profits, relying if necessary on side payments. If all banks share the identical subjective evaluations of the profitability of borrowers' investment projects, then clearly the optimum rate R_i^* to be charged to the ith customer would be the same no matter which bank served him. Furthermore, even allowing for differences in the subjective evaluation of borrower risk and assuming an arbitrary initial distribution of customers between banks, the device of buying and selling customers would allow each bank and the industry as a whole to maximize profits. In this way a banker's Pareto optimum would be reached with each bank charging its customers the monopolist rate, and thus, again, no credit rationing would occur.

In this section we propose to argue that

this solution is in fact not feasible, at least in the present American economy. We suggest, instead, that banks can best exploit their market power, while remaining within the bounds set by prevailing institutions, by classifying customers into a rather small number of classes within each of which a uniform rate is charged, even though the membership of each class will exhibit considerable heterogeneity in terms of R_i^*.

First, even if there were but a single monopoly bank or a perfectly collusive banking system, the mere existence of usury laws would lead toward the indicated solution. Such laws would prevent the banker from charging any rate R_i^* which is greater than the legal limit. Thus all customers for whom R_i^* is larger than the ceiling would be classified together in the category with the ceiling rate. Since the monopolist rate R_i^* for each customer in this class would equal or exceed the uniform rate set for the class, namely the usury ceiling, it is apparent from the results of Section I that many, if not all, customers in this class would be profitably rationed.

Even aside from usury ceilings, the pressure of legal restrictions and considerations of good will and social mores would make it inadvisable if not impossible for the banker to charge widely different rates to different customers. A banker would tend, instead, to limit the spread between the rates and to justify the remaining differentials in terms of a few objective and verifiable criteria such as industry class, asset size, and other standard financial measures. An effort would no doubt be made to choose the criteria for classification so as to minimize the difference between the optimal classification of customers into rate classes and the categories dictated by the objective criteria, but a close approximation might be difficult to achieve.

The inducement to adopt a classification scheme of the type described is likely to be greatly strengthened when we take into account the fact that banks cannot openly collude, although they share a common desire to maintain rates as close as feasible to the collusive optimum. In order to prevent, or at least minimize, competitive underbidding of rates they would need tacit agreement as to the appropriate rate structure for customers, and thus a classification scheme based on readily verifiable objective criteria would appear as an efficient and effective device. Furthermore, to make the whole arrangement manageable, the number of different rate classes would have to be reasonably small. Finally, one can also readily understand how such tacit agreement on the structure of class rates could be facilitated by tying these rates through fairly rigid differentials, to a prime rate set through price leadership.

If we now superimpose the impact of usury ceilings along with the other legal and social constraints, it is clear that the entire structure of rates would tend to be compressed within narrower limits than would otherwise be optimal. This means, in particular, that the rate for each class would tend to the lower limit of the R_i^* spread appropriate for the customers in that class, with the possible exception of the lowest class rate reserved for the riskless or nearly riskless prime customers. The result is that widespread rationing would occur, particularly in the higher rate classes.

Finally, we may observe that the oligopolistic price setting pattern outlined above is likely to lead to a very sluggish and somewhat jerky adjustment of the entire rate structure as changes in underlying conditions generate changes in the optimal level and structure of rates. The considerations relevant here are well known from the literature on oligopolistic

market structure and price leadership. It follows that when conditions are changing rapidly we might expect to find that the entire structure of rates would lag behind, and thus, for a while, would tend to be higher or lower than the optimal structure, depending on the direction of the change. These considerations provide the key to dynamic rationing to be elaborated below.

2. *Long-Run Equilibrium Credit Rationing*

We are now in a position to assemble the complete theory of credit rationing. We first take up equilibrium credit rationing which occurs when the loan rate is set at its optimal level. In Section I it was shown that credit rationing will be profitable, even in long-run equilibrium, as long as there is uncertainty of loan repayment *and* banks cannot discriminate perfectly between customers. In Section II.1 both of these conditions were verified as features of the commercial banking industry, and thus we conclude that equilibrium rationing is consistent with rational economic behavior.

This conclusion would remain basically unchanged if one recognizes that the bank's expected return and cost is affected by other contract terms such as loan maturity and compensating balance requirements. To be sure, if a bank could discriminate freely between customers with respect to such factors, there might be no occasion for equilibrium rationing. But while the empirical evidence is scanty, one would surmise that banks are limited in their power to discriminate with respect to these terms as well as the nominal interest charge. Similarly, our analysis can incorporate the benefits of maintaining long-run "customer relationships" (cf. Hodgman [4], Edward Kane and Burton Malkiel [6]) without invalidating our basic conclusions about the rationality of rationing.[16]

It is worthwhile to consider briefly the comparative static properties of equilibrium rationing in our model. The total amount of equilibrium rationing, E, can be written as the difference between loan demand and loan supply for those customers of the bank experiencing rationing:

$$E = \sum_{i=1}^{n} \max\,[D_i(\hat{R}) - L_i(\hat{R}_i), 0] \tag{8}$$

The excess supply of loans to those customers to whom the bank would like to extend additional loans does not, of course, offset the rationing to other customers. The offer curve to a firm depends on the opportunity cost of the bank, I, and the firm's density function of possible outcomes, f_i, and hence E depends on these two factors and the firms' demand functions. Consequently, for purposes of comparative static analysis, one may consider the impact of changes in demand, risk, and opportunity cost on the amount of equilibrium rationing. Unfortunately, the evaluation of the impact of these changes turns out to be a very difficult process. In particular, the rationed or nonrationed status of customers may change because of the change in the underlying parameter. To at least illustrate the possible outcomes, however, we shall outline the case of a change in the opportunity cost I.

Two important results for the comparative static analysis are given in the following propositions:

PROPOSITION 8. A *ceteris paribus* increase in the opportunity cost causes the optimal loan offer curve to shift downward at every interest rate. Indeed, by implicit differentiation of equation

[16] Although the customer relationship and nonprice contract terms may serve to modify the importance of uncertainty in our theory of credit rationing, the bank's inability to discriminate perfectly in setting these terms is still critical. The need for such an element can be seen clearly in the attempts by Hodgman [4, p. 265] and Kane and Malkiel [6, p. 123] to rationalize credit rationing without appeal to such imperfect discrimination.

(3) we obtain:

$$\frac{\partial L_i}{\partial I} = \frac{-1}{f_i R^2} < 0 \tag{9}$$

PROPOSITION 9. The common optimal rate factor R^* (applicable if the bank were constrained to satisfy all demand functions) is positively related to the opportunity cost I.

Proposition 8 is self-evident. Proposition 9 is in line with the standard theory of the firm in that an increase in marginal cost necessitates an increase in marginal revenue which takes the form of an increase in R^*. Because the very act of changing the parameter I may lead to a change in the rationed or nonrationed status of customers, as demonstrated below, special cases can arise in which $\hat{R}$ (the common optimal rate allowing for rationing when profitable) and I are not positively related, contrary to the relationship between R^* and I. Since such cases are distinctly abnormal, the obvious and reasonable situation being a positive relationship between I and $\hat{R}$, we shall proceed on the premise that proposition 9, is equally applicable to $\hat{R}$.[17]

With the aid of Figure 2, we can survey the impact of an increase in the opportunity cost from I_0 to I_1. From proposition 8, it is known that the offer curve will shift downward from the locus $\hat{L}_0$ to $\hat{L}_1$; and from proposition 9 it is known that the common optimal rate factor will increase, say from $\hat{R}_0$ to $\hat{R}_1$. Thus the net impact of a change in the opportunity cost on the maximum loan offered is the result of a shift in the offer curve and a movement along this curve, and the amount of equilibrium rationing may either rise or fall. These two alternative outcomes are illustrated in Figure 2. In the case of the demand curve D_1, the customer would move from a rationed to a nonrationed status, while for the case illustrated by the demand curve D_2, the opposite conclusion holds. Cases in which the rationing status remains unchanged are, of course, equally possible. Thus the effect of an increase in I on the size of rationing must remain uncertain. This result, however, should not be regarded as disturbing. What it implies is that in the long run, ***no systematic relationship exists between the extent of equilibrium rationing and the absolute level of interest rates.***

3. *Dynamic Credit Rationing*

We define dynamic rationing as the difference between equilibrium rationing and the volume of rationing that arises when the actual rate charged customers, R, differs from the long-run equilibrium rate, $\hat{R}$. By this definition, dynamic rationing can be positive or negative, and we shall show that its magnitude will be positively associated with the spread, $\hat{R}$-R. Furthermore, we have already suggested, in view of the oligopolistic structure of the banking industry, that R is likely to adjust slowly to changes in $\hat{R}$, thus lending

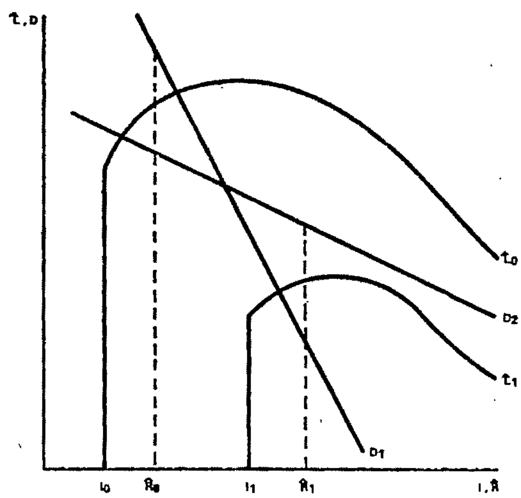


FIGURE 2

[17] The abnormal conditions which lead to this abnormal result are discussed further in [5, p. 56 and p. 86].

credence to the empirical importance of dynamic credit rationing. In fact, there is evidence that banks may tend to rely on some objective signal, such as changes in the Federal Reserve's discount rate, in determining the timing of loan rate changes, and hence, dynamic credit rationing would vary significantly depending upon Federal Reserve policy.[18]

It is helpful to begin by considering a system that starts in long-run equilibrium with a rate, $\hat{R}_0$, and then receives a shock that changes $\hat{R}_0$ to $\hat{R}_1$ while the quoted rate remains unchanged at $\hat{R}_0$. As we have seen, there are three main market forces that can change $\hat{R}$: a change in market interest rates leading to a change in the opportunity cost I; a shift in customer demand schedules; and a change in risk as may be indicated by a shift in $F[x]$. We shall first consider the effects of changing each of these factors, one at a time. Then by combining the results of these *ceteris paribus* experiments, we can examine more realistic situations in which several factors change simultaneously.

1. Consider first a change in market rates of return which, for convenience, may be summarized by some representative rate, r_M. a rise in r_M will directly increase the opportunity rate for funds invested in the loan portfolio. From propositions 8 and 9, we know that such an increase in I will lower all offer curves and also increase $\hat{R}$. Since the quoted rate remains at $\hat{R}_0$ and all the demand curves are unchanged, *there must be an aggregate increase in rationing*. Our model also provides some information about the incidence of this increase. At one extreme, the risk free firms will still not experience any rationing because the offer curve for them is a vertical line.[19] At the other extreme, firms already rationed in the initial equilibrium will be rationed even more as the offer curve shifts down. Finally, among the risky firms initially not rationed, some will remain unrationed while others will experience new rationing, with the amount depending upon the extent of the shift.

Thus on the whole, the loan portfolio will shrink and the funds released by rationing will be shifted into other assets whose yield has increased (or be used to repay the now more costly borrowed funds). However, the opportunity cost I is now likely to fall relative to the market rate r_M because loans will be a smaller percentage of the total portfolio. This will tend to moderate, though not eliminate, the initial increase in rationing.

2. Consider next the effect of a downward revision of the anticipated distribution of outcomes; operationally, we may think of the initial distribution $F[x]$ being replaced by a new one $G[x]$ with $G[x] \leq F[x]$. As can be verified from (3), such a re-

[18] Empirical tests performed in an earlier work by one of the authors [5, Ch. 4] confirms the hypothesis that, in the period covered by our data, changes in the Federal Reserve discount rate were a major factor influencing the timing of changes in the commercial loan rate. It should be stressed, though, that such a relationship might not continue to hold in the future unless the Federal Reserve continues to operate the discount window in the customary fashion. Should the recent Federal Reserve proposal [11] to keep the discount rate closely in line with market rates be enacted, it is likely that the commercial loan rate would adjust more quickly toward its desired level. In this case, dynamic rationing in response to changes in market conditions would tend to die out faster than in the past. Thus discount rate policy is at least one way in which the Federal Reserve can influence the amount of credit rationing.

[19] This conclusion would not hold if the shift in I were so large as to exceed $\hat{R}_0$. In this case, the best course of action would be to cut off loans even to the prime firms, and a fortiori to all customers. But this case can be disregarded for the banks could be counted upon to respond promptly by raising their quoted rate, i.e., $\hat{R}_0$ would not remain unchanged in these circumstances. More generally in the usual theory of monopoly or oligopoly, if the equilibrium price rose because of a shift in either the demand or the cost function, and, for some reason, the market price was prevented from rising, rationing would occur only if the shift were such that the marginal cost would exceed the price. It appears, therefore, that dynamic rationing, just as equilibrium rationing, is intimately related to the uncertainty about the outcome of the loan.

vision results again in a downward shift in the offer curve, and in an increase in the R_i^* and hence in $\hat{R}$. Since the quoted rate has not changed, the maximum loan offered at that rate must tend to decline. With the demand unchanged, rationing must therefore increase for some of the customers, and the incidence of the increased rationing is entirely analogous to that of case 1. Once more, as loans shift out of the now less remunerative loan portfolio, the opportunity cost I may decline somewhat, mitigating the initial effects.

3. Much the same conclusion can be seen to hold for the case of a shift in all demand schedules, r_M and $F[x]$ constant, except that in this case, it is the demand curve which shifts while the offer curve remains unchanged. If I remains unchanged, then the initially rationed customers will experience more rationing because their increased demand is not satisfied at all; the risk free customers, in contrast, will receive larger loans and remain unrationed; and the initially unrationed risky customers will also obtain larger loans, but possibly not enough to match the increase in their demand. Furthermore, as funds flow into the loan portfolio (because some of the increased demand is satisfied), the opportunity cost of loans will tend to increase giving rise to additional rationing of the type under (1).

Normally, an increase in loan demand will tend to occur in periods of buoyant economic activity and hence will be associated with a rise in r_M (partly reflecting the higher demand in all markets and partly causing, in turn, a higher demand for bank funds), and also with an increase in the anticipated profitability of the investment projects, the latter producing a decrease in risk as measured by $F[x]$. The outcome is then a combination of the *ceteris paribus* results under (1), (2), and (3). The shift in $F[x]$ tends to raise the loan offered at the unchanged rate, $\hat{R}_0$; at the same time, the quantity demanded at that rate rises, and if I were unchanged, the effect on $\hat{R}$, as well as on rationing, would depend on the relative amount of the two shifts. Even on this basis, one would anticipate an increase in rationing because the demand curve is likely to shift further than the offer curve, reflecting an increase in the optimism of firms relative to that of the banks. But more fundamentally, I must rise, both because with I constant, more funds would flow into the loan portfolio causing I to rise relative to r_M, and because r_M itself will be rising. Hence the final outcome will tend to be the same as in the previous three cases. The extent of increased rationing will depend on the relative shift in $F[x]$, in the quantity demanded at the unchanged rate, and in r_M, and on whether, on balance, these shifts will cause funds to flow in or out of the loan portfolio. It is apparent, however, that the rise in I and in rationing will tend to be greater the smaller the elasticity of supply of funds to the banking system. The effect might be particularly severe if the ability of banks to attract funds were actually reduced, as happened in some recent episodes in which the Certificates of Deposit rates in secondary markets pierced the Regulation Q ceiling.

We may thus conclude quite generally that as $\hat{R}$ rises relative to $\hat{R}_0$, rationing will tend to increase; and the incidence of the increased rationing will tend to fall most heavily on customers who would be rationed in equilibrium, and will tend to affect the least, if at all, the riskless, or nearly riskless, prime customers.[20] One implication of this

[20] As we have argued in the text, the amount of dynamic rationing tends to be positively related to the spread between the equilibrium rate and the quoted rate. Actually, though this relationship must hold in the large, in some special circumstances it may fail to hold in the small. The special cases arise only when the slope of the demand curve exceeds the slope of the offer

result is worth stressing since it provides the key to our operational measurement of credit rationing set forth in the next section. Suppose that we were to classify all customers into two broad classes, the prime customers and all others. We should then expect that as the gap between $\hat{R}$ and R widens and dynamic rationing becomes more severe, loans to the riskless customers will tend to represent a growing share of the total loan portfolio.

This result can be given the following useful interpretation, which also serves to bring to light the common nature of equilibrium and dynamic rationing. In the presence of risk as to the outcome of the loan, reducing the size of the loan will increase the expected rate of return, by reducing the expected loss from insolvency of the firm. It is therefore quite understandable that a bank faced with a higher opportunity cost (whether from a rise in the market rate or in lending opportunities) and unable to raise the return by raising rates, will find it profitable to raise its return at least by upgrading the quality of its portfolio through a reduction in risk; the upgrading may take the form of shifting funds toward less risky customers, and/or of reducing loans made to risky customers, depending on the nature of the shift in underlying conditions.

III. *A Measure of Credit Rationing*

In this section we shall develop an operational measure of credit rationing which is based on the theory as developed in Sections I and II and which is used in the test of the theory presented in Section IV below. In principle, the volume of credit rationing should be measured by the difference between the loan demand and bank supply for rationed customers as

curve. Since this can occur only at relatively high interest rates, indicating default risks well above levels bankers would consider acceptable, the case would seem to have little empirical relevance.

defined by E in equation (8). The degree or relative incidence of credit rationing could then be measured by the ratio of the volume of rationing to the potential demand of rationed customers, or:

$$\hat{H} = \frac{E}{E + L_2} = \frac{D_2 - L_2}{D_2} \tag{10}$$

where D_2 denotes the demand of rationed customers and L_2 the volume of loans actually granted to them.

Unfortunately, the direct measurement of E and L_2, the components of $\hat{H}$, requires information on the *ex ante* customer demand and bank supply which is unlikely to be available, even in the future. The analysis of Section II.3, however, points to a possible, operational, proxy measure of the degree of dynamic credit rationing. As shown there, our model suggests that there should be a positive association between variations in dynamic credit rationing and variations in the proportion of the total loan portfolio accounted for by the risk free prime customers. Let us then denote by $L_1 \simeq D_1$, the volume of loans granted to these customers, and by L_2 and D_2, respectively, the loan granted and the loan demanded by all other customers. Our proposed operational proxy for the non-observable $\hat{H}$ is either of the following two:

$$H_1 = \frac{L_1}{L_1 + L_2} \tag{11.a}$$

or

$$H_2 = \frac{1}{H_1} = \frac{L_1 + L_2}{L_1} \tag{11.b}$$

The first measure, H_1, is simply the percentage of total loans which are granted to the risk free customers, and hence is positively related to the degree of rationing. The alternative proxy is its reciprocal and hence, is negatively related to the degree of rationing.

To see the relation between either proxy

and the ideal measure $\hat{H}$, let

$$B = \frac{D_2}{D_1}$$

Then from (11) we obtain:

$$H_1 = \frac{1}{[1 + B(1 - \hat{H})]} \tag{12.a}$$

$$H_2 = 1 + B(1 - \hat{H}) \tag{12.b}$$

since $L_1 = D_1$.

Differentiating (12) with respect to $\hat{H}$, yields:

$$\frac{\partial H_1}{\partial \hat{H}} = \frac{B}{[1 + B(1 - \hat{H})]^2} > 0 \tag{13.a}$$

$$\frac{\partial H_2}{\partial \hat{H}} = -B < 0 \tag{13.b}$$

The equations (12) show that for a given value of B, our proxies are monotonic functions of the ideal measure $\hat{H}$, and (13) confirms that the relation is in the direction expected. The functions relating either proxy to $\hat{H}$ involve as a parameter the relative demand factor, B, essentially because we have replaced the nonobservable D_2 with the observable D_1. This of course implies that any change in B will give rise to a variation in H_1 or H_2 which does not correspond to variations in $\hat{H}$. Hence, if there were sizable variation in B over time, our proxy measures of $\hat{H}$ could be subject to appreciable errors in measurement. Note, however, that even in this event, as long as H_1 or H_2 are used as dependent variables in a statistical test as we shall do below, these errors will not tend to generate bias in the estimated coefficients unless B happens to correlate with the behavioral determinants of $\hat{H}$.[11] Fortunately, the conclusions of Section II concerning the comparative static properties of the model and the classification of customers suggest that, at least in principle, these variables would not be correlated.

Some data problems are encountered even in attempting to measure the proxy developed in equations (11). Our source of data is the Federal Reserve's "Quarterly Interest Rate Survey" which records the volume of new loans granted by rate and size class during the first two weeks of the last month of the quarter. With this data, the percentage of loans granted to risk free firms (proxy H_1) can be at least approximated by the percentage of loans granted at the prime rate (and perhaps secondarily by the percentage of loans which were large in size). Unfortunately, in several instances the prime rate changed during the period of the survey, with the effect that one cannot distinguish between loans made at the new prime rate and loans made at this same rate while the old prime rate was still in effect. To circumvent this problem and secure a more reliable measure, we smoothed these quarters as well as possible. Following John Hand [2] who first used the data for this purpose, we combined the smoothed series with three other measures based on the distribution of loans by size through principal components analysis. More specifically, we calculated the first principal component of the following four series:

a) The proportion of total loans granted at the prime rate.
b) The proportion of total loans over \$200,000 in size.
c) The proportion of loans over \$200,000 in size granted at the prime rate.
d) The proportion of total loans \$1,000–\$10,000 in size.

The first three series should enter positively into the principal component and

[11] Because (12a) is nonlinear in $\hat{H}$, this statement will be true only for a linear approximation of H_1. If either proxy were being used as an independent variable, the problem of bias would be more serious. See, for example, Malinvaud [7, Ch. 10].

the fourth negatively, of course. The principal component has a mean of zero and standard deviation of unity and the four factor loadings were, respectively:

(a) .988

(b) .968

(c) .939

(d) −.959

This indicates that each series enters prominently and about equally into the principal component.

The principal component thus derived corresponds to the H_1 measure since the series (a) is analagous to H_1. Exploratory calculations indicated that the results would not be significantly affected by relying on the alternative H_2 measure derived from the reciprocal of series (a) to (d).[22] The solid line in Figure 3 is a plot of the seasonally adjusted principal component H to be used in the tests of the following section. The pattern of rationing indicated by the proxy seems quite credible throughout the period. Note in particular how the most recent pattern is consistent with what we might have expected, rising very high in the second and third quarters of 1966 and then falling off somewhat in the fourth quarter. Unfortunately this series cannot be computed beyond 1966 at the present time, since the information produced by the loan survey beginning in the first quarter of 1967 is not strictly comparable with the earlier information.[23]

IV. *A Test of the Model Using the Credit Rationing Proxy*

In this section we propose to use the credit rationing proxy to test the implications of our model as to the forces controlling variations in time in the extent of credit rationing. This test, relying on quarterly time series data for the years 1952 to 1965, will thus serve to shed light on three aspects of our problem: our theory of credit rationing, the effectiveness of the proxy variable based on the theory, and the existence of rationing as an empirically significant phenomenon.[24]

The principal implication of our theoretical model is that the main source of systematic variations in credit rationing is to be found in changes in dynamic rationing; and that these changes in turn are positively associated with the spread between the long-run optimal or equilibrium loan rate denoted hereafter by r_L^*, and the rate actually prevailing, r_L.[25] If we further assume that, to a first approximation, this association can be formulated as a linear relationship within the empirically relevant range, we are led to

$$H = a_0 + a_1(r_L^* - r_L) + \epsilon \tag{14}$$

Recall that, according to our model, when the commercial loan rate is at its long-run desired level, so that the second term of (14) is zero, we have only equilibrium

[22] We estimated a number of equations using the rationing model developed in Section IV for the specification of the independent variables, and the individual series (a) to (d), the principal component, and the reciprocal of each as separate dependent variables. The principal component yielded the best fit as expected, but comparable results were obtained with the other measures. We are grateful to John Hand for making his data on these series readily available.

[23] The difficulty arises from the fact that with the first quarter of 1967, the period of the survey was changed from the first two weeks of the last month of each quarter to the first two weeks of the second month of each quarter. Because of the strong seasonal component in the four series used to compute H, sufficient observations for the new survey period must be obtained before the new seasonal component can be reliably determined.

[24] It is important to note that success in this test will confirm the value of the proxy as a variable to be used in testing for the impact of credit rationing on the real sectors of the economy. Indeed, the first uses of the proxy for this purpose have been made in [2] and [5].

[25] r_L^* denotes an empirical approximation to the theoretical construct $\hat{R}$ developed above. In particular, r_L^* must stand proxy for the spectrum of optimal rates corresponding to the respective risk classes. Similarly, r_L stands for the spectrum of actual rates and is measured as the average rate on commercial loans compiled from the "Quarterly Interest Rate Survey."

$$r_L^* = c_0 + c_1\{r_T + b_1[(DEP/TB) - 1] - b_2D62 + b_2'C + b_3L/(A - L) + b_4\Delta[L/(A - L)]\} \tag{15}$$

rationing. On the microeconomic level, equilibrium rationing depends on the specific parameter values for the demand functions, density functions, and opportunity cost. But our investigation in Section II.2 indicated that no systematic relationship existed between the degree of rationing and changes in these parameters. Accordingly in (14), the constant term a_0 may be thought of as a measure of equilibrium rationing up to a stochastic error term which is included in the overall error term ϵ. When the quoted rate is above the equilibrium rate, the amount of dynamic rationing can be considered negative in the sense that rationing will be reduced below its equilibrium level, even though the actual amount of rationing can never be less than zero by definition. The dependent variable as measured by our principal component proxy will, in fact, take on negative values because the zero point for H is chosen arbitrarily. The arbitrary choice of origin is, of course, reflected in the constant a_0, and consequently one cannot distinguish between the level of equilibrium rationing and the scaling effect in the constant.[26]

We now turn to the important task of specifying the equilibrium commercial loan rate, r_L^*. Our theory indicated that the desired commercial loan rate would be at that level at which the marginal proceeds from a commercial loan, after adjustment for risk, would just equal the opportunity cost. We also know that this relationship will be valid for all other assets in the bank's portfolio. This means that the optimal commercial loan rate will tend to equal the market yield on any other asset held in the bank's portfolio after adjustment for risk, maturity, liquidity, and, possibly, any expectations concerning future levels of that rate. We are free, then, to choose as the standard of comparison, any security which is widely held by banks, the obvious criterion being practical expedience. Our choice, on this basis, is the bank's holdings of Treasury bills.

On the basis of the above considerations we are led to equation (15) as the specification for the desired commercial loan rate, where the notation will be defined as we proceed.

Consider first the bank's return on Treasury bills. It may be regarded as consisting essentially of two components. The first component, the Treasury bill rate, r_T, is straightforward. The second component is the liquidity value of Treasury bills, which is more involved and, in fact, accounts for the second, third, and fourth terms in the brackets. Our basic premise is that the liquidity yield of Treasury bills should decrease as the bank's holdings of these bills (TB) rises relative to its deposit liabilities (DEP). Our specification for this liquidity term takes the form b_1 [(DEP/TB)-1] where b_1 (>0) is an estimated parameter.[27]

[26] Equation (14) can be rewritten such that the spread between the desired commercial loan rate and the rate actually quoted is a linear function of the degree of credit rationing. In addition, it has already been suggested that the timing of adjustments in the commercial loan rate depends on changes in the Federal Reserve discount rate. These two factors can be combined into a testable partial adjustment model of the determinants of the commercial loan rate in which the size of the desired adjustment depends on the degree of rationing while the speed of adjustment depends on changes in the discount rate. See fn. 18.

[27] The legal requirement that banks must maintain government securities in their portfolio as collateral for government deposits raises one conceptual problem. To the extent that this requirement necessitates holding

The liquidity value of Treasury bills has been reduced since about 1962, however, by the development of a broad and active market in Certificates of Deposits (*CD*'s), which affords the banks an important means for increasing their liquidity at short notice. The effect of the existence of this market may be measured by a simple shift parameter or dummy variable and this is included as the term $-b_2\ (D62)$ where $D62$ is a dummy variable which is unity starting in 1962-I and zero before then.[28] The value of the *CD* market is of course severely limited when the Regulation Q ceiling on *CD* interest rates is binding. This countervailing effect may be specified by an additional dummy variable which takes the value one in those quarters, if any, in which the secondary market rate for *CD*'s exceeds the ceiling rate. One would expect this dummy variable, denoted by C, to have a coefficient opposite in sign and of the same order of magnitude as the *CD* dummy variable $D62$.

The next to last term in equation (15) measures the share of the loan portfolio in total assets which, as suggested in Section II, should tend to affect the opportunity cost of funds for loans relative to market rates. In addition, this variable may be visualized as an adjustment of the required rate on loans for their relative illiquidity. We anticipate that this illiquidity should increase at an increasing rate as the ratio of loans to assets (or liabilities) grows and thus measure this effect as $b_3(L/(A\text{-}L))$ where $b_3\ (>0)$ is an estimated coefficient, A is total loans and investments, and L is the commercial loan portfolio of the banks. Finally, we should also consider changes in the liquidity ratio, since a dynamic short-run increase in loans which is beyond the control of the bank would have additional (although only transitional) liquidity cost, and this effect is accounted for by the last term in (15).

We have now almost completed the task of specifying the desired commercial loan rate. To obtain more generality we have formulated the desired loan rate as a linear function of the terms just summarized. The need for the linear function arises because we have not yet formally accounted for differences in risk and maturity between commercial loans and Treasury bills. The basic equation to be estimated can now be derived by substituting equation (15) into (14), which yields:

$$
\begin{aligned}
H = {} & (a_0 + a_1c_0 - d_1b_1) - a_1(r_L) \\
& + d_1(r_T) + d_1b_1(DEP/TB) \\
& - d_1b_2(D62) + d_1b_2' C \\
& + d_1b_3[L/(A - L)] \\
& + d_1b_4\Delta[L/(A - L)]
\end{aligned}
\tag{16}
$$

where $d_1 = a_1c_1$.

The coefficients of (16) were estimated using ordinary least squares from observations for the period 1952-II to 1965-IV.[29] (The year 1966 was omitted to enable us to carry out extrapolation tests reported below.) The results are as follows:

Treasury bills of some required amount, the correct variable for our analysis would be the free bills; that is, the bills held above the required amount. It has been suggested that, at least in 1966, such a restriction was a restraining influence on bill holdings. Legally, however, a wide variety of Federal and Local Government securities are acceptable as collateral and it is unlikely that more than a few banks, probably centered in New York City, held Treasury bills only to satisfy collateral requirements, even in 1966.

[28] We have also experimented with incorporating the *CD* term in the form of a multiplicative factor operating on the Treasury bill ratio itself. Since the two estimates are nearly identical, only the results for the linear form will be shown.

[29] The source of data for the independent variables is the *Federal Reserve Bulletin* with the exception of commercial loans. Commercial loans are the sum of industrial and commercial loans (from an unpublished Federal Reserve series) and nonresidential mortgage loans of the commercial banks (from the *Federal Reserve Bulletin*). The mortgage loans are included since they are made to the same customers as the shorter maturity commercial loans. All dollar magnitudes are seasonally adjusted and interest rates are measured as a percent.

$$(17)\quad H = \underset{(-7.0)}{-3.48} - \underset{(-5.3)}{1.16r_L} + \underset{(2.2)}{.296r_T}$$

$$+ \underset{(1.2)}{.007DEP/TB} - \underset{(-7.8)}{1.27D62}$$

$$+ \underset{(1.6)}{.614C} + \underset{(8.0)}{26.7}[L/(A-L)]$$

$$+ \underset{(.92)}{9.0}\Delta[L/(A-L)]$$

$S_e = .350 \qquad R^2 = .84 \qquad D.W. = 2.02$
(T statistics shown in parentheses)

These estimates confirm the implications of our model in that all coefficients have the correct sign and the goodness of fit is respectable, taking into account the noise in the dependent variable which was discussed above. Note in particular the very significant negative coefficient of the commercial loan rate. This supports one of the most distinctive implications of our model, to wit, that a rise in the commercial loan rate given the optimal level of this rate as measured by the remaining variables in the equation, tends to reduce rationing as it reduces the demand and increases the supply. Similarly, the opportunity cost as measured by the Treasury bill rate, and the loan illiquidity variable, measured by the share of loans in the bank's portfolio, appear as the most significant factors tending to increase rationing, given the commercial loan rate. Although the Treasury bill liquidity term and the change in the commercial loan illiquidity term are not significant, this is probably due to multicollinearity since in the short run, with the level of total assets essentially fixed, banks may have to sell Treasury bills to meet unexpected loan demand.

The remaining variables are the dummies intended to measure the effect of *CD*'s. The coefficient of the *CD* dummy, *D*62, is both very significant and large (since the dependent variable has by construction unit variance). It suggests that the newly acquired ability of banks to attract and shed funds through *CD*'s contributed appreciably to a reduction of rationing. The contribution of the ceiling rate dummy *C*, on the other hand, is harder to assess because the ceiling was binding, and hence the dummy was one, only in the last quarter of the sample, 1965-IV.[30] Its coefficient has the expected positive sign but its magnitude, which we had expected to be roughly equal to that of the coefficient of *D*62, is instead only half as large. Yet this result makes good sense when we recall that the ceiling rate was effective only to December 6, which is just about in the middle of the two week period during which the loan survey is taken. Indeed, it suggests that if the ceiling had been effective throughout the period, then its effect would have come close to offsetting totally the contribution of the *D*62 dummy as expected.

The effectiveness of our model in explaining the behavior of the rationing proxy can also be judged from the plot of the values of *H* computed from equation (17) which is shown by the dotted line in Figure 3. The equation appears to track the broad movements of the actual series as well as the major turning points with leads or lags not exceeding one quarter. The largest errors correspond to a number

[30] The multiplicity of *CD* maturities and the thinness of the secondary market make it difficult to obtain a reliable indicator of those periods in which the ceiling is binding. In addition, an aggregation problem arises because the ceiling may be binding only for some banks or for some regions. The available information suggests, though, that until the last quarter of 1965 the ceiling rate was promptly raised whenever it threatened to become a significant hindrance to the issuing banks. In October and November of 1965, however, the secondary rate rose unequivocably above the ceiling and remained there until December 6th when the ceiling was again raised. See Willis [10] for further discussion of the secondary market for *CD*'s.

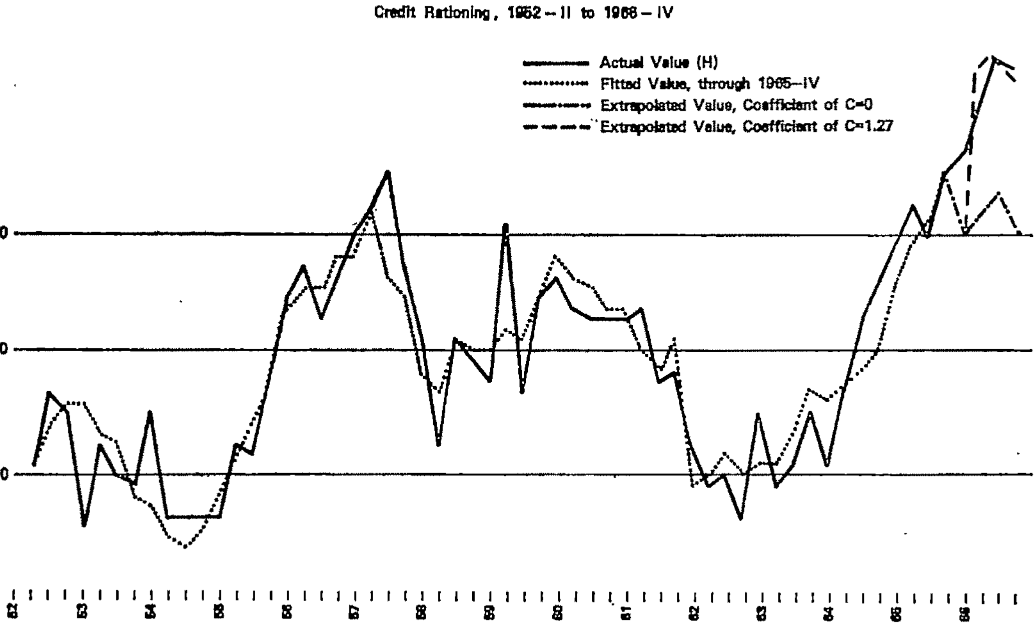


FIGURE 3

of one quarter spikes in H which may well reflect mostly noise in that series.

In Figure 3 we also present some extrapolations of our equation to the year 1966, which marks the end of the period for which usable data on the rationing proxy are presently available.[31] Unfortunately extrapolations of (17) to 1966 run into rather formidable difficulties because throughout the last three quarters the ceiling rate on CD's fell short of the secondary market rate. Nonetheless we feel it worthwhile to exhibit these extrapolations because of the tentative light they shed on the working of the ceiling rate. If one extrapolates (17) as though the ceiling rate dummy C were zero, one obtains the computed values represented by the dotted dashed line and as expected, this extrapolation very much underestimates the extent of rationing in the last three quarters. In order to allow for the ceiling effect, an alternative extrapolation of (17) was carried out for the last three quarters, represented in Figure 3 by the dashed line, in which the dummy variable C was assigned the value of one. In addition, since the ceiling rate was effective throughout these quarters, it was also assumed that the coefficient of C was equal numerically to that of $D62$. Stated differently, the alternative extrapolation assumes that a binding ceiling has the effect of undoing the loosening effect of CD's measured by the dummy $D62$. It is seen that this alternative fits the observations remarkably well.

These results, if taken at face value, have rather interesting implications for the *modus operandi* of ceiling rates as a tool of monetary policy. In particular they would support the view that, by allowing the ceiling rate to become a real hindrance to the ability of banks to attract new CD funds, the Federal Reserve could reduce significantly the availability of funds to the commercial bank customers of the banking system. This reduction would presumably occur to the benefit of customers of other intermediaries and/or of

[31] Cf. fn. 23.

those firms able to raise funds directly in the market. It should be recognized, however, that our evidence in support of this inference is at the moment rather limited and hence, quite tentative, until it can be confirmed by further experience under similar circumstances. Of course, by the time the system is again exposed to similar circumstances, it may have learned ways of evading or by-passing, at least partially, the constraint imposed by the ceiling.

REFERENCES

1. M. Freimer and M. J. Gordon, "Why Bankers Ration Credit," *Quart. J. Econ*, Aug. 1965, *79*, 397–416.
2. J. Hand, *The Availability of Credit and Corporate Investment*, unpublished doctoral dissertation, M.I.T. 1968.
3. D. R. Hodgman, "Credit Risk and Credit Rationing," *Quart. J. Econ*, May 1960, *74*, 258–78.
4. ———, "The Deposit Relationship and Commercial Bank Investment Behavior," *Rev. Econ. Statist*, Aug. 1961, *43*, 257–68.
5. D. M. Jaffee, *Credit Rationing and the Commercial Loan Market*, unpublished doctoral dissertation, M.I.T. 1968.
6. E. J. Kane and B. G. Malkiel, "Bank Portfolio Allocation, Deposit Variability, and the Availability Doctrine," *Quart. J. Econ.*, Feb. 1965, *79*, 113–34.
7. E. Malinvaud, *Statistical Methods of Econometrics*, Chicago 1966.
8. M. H. Miller, "Credit Risk and Credit Rationing: Further Comment," *Quart. J. Econ.*, Aug. 1962, *76*, 480–88.
9. R. Roosa, "Interest Rates and the Central Bank," *Money, Trade, and Economic Growth in Honor of John Henry Williams*, New York 1951.
10. P. B. Willis, "The Secondary Market for Negotiable Certificates of Deposits," paper prepared for the Committee for the Fundamental Reappraisal of the Discount Mechanism, appointed by the Board of Governors of the Federal Reserve System.
11. Board of Governors of the Federal Reserve System, *Reappraisal of the Federal Reserve Discount Mechanism*, Washington 1968.

Pecuniary and Technological Externality, Factor Rents, and Social Costs

By Dean A. Worcester, Jr.*

Analysis of external effects in production has turned during the last 15 years from analysis of pecuniary and technological economies and diseconomies of well defined industries (and a belief that the effects are minor),[1] to analysis of interdependence among very few, often only two, individual utility or production functions (and the belief that the effects may be important).[2] This shift seems to have been inspired by Ronald Coase [4]. He is the first to have shown that firms in competitive industries which have interdependent production functions will find negotiations that properly adjust for the externality to be profitable. He also shows that merger of the two firms will result in optimal resource allocation.

No one has argued recently that all of the externalities in even a small enterprise economy can be overcome by negotiation, merger of the interdependent firms, or by wholesale merger of all firms into an administratively planned economy. On the contrary, recent work has emphasized the difficulties that make even conceptual solutions appear virtually impossible for a wide range of cases.

Neoclassical analysis, which rests on the relationships between firms and industries, is utilized in this paper for five reasons. (1) It is adequate to reveal the basic source of externalities. (2) It provides a logical progression of steps for the extension of the analysis. (3) It is more readily applicable to feasible social controls that can, in many cases, reduce the considerable distortions due to the presence of externalities. (4) It lends itself to an analysis of the welfare effects of alternative market structures in the presence of externalities. (5) It is appropriate for both separable and nonseparable cost functions as defined by Otto Davis and Andrew Whinston [6, p. 245].

The principal points to be made here are: (1) Technological economies and diseconomies occur when the firms comprising the industry are not able to equalize the ratios of marginal products to factor price ratios which reflect alternative costs, because one or more of the factor costs to the firm do not equal opportunity cost to the industry (or to society). This is often the case, arising when the technical production function of an industry is not accommodated properly by the prevailing system of ownership, legal rights and the like. Technological diseconomies exist when, in

* Department of economics, University of Washington. He wishes to thank José Encarnacion and J. Timothy Peterson for their helpful comments on an earlier draft and the anonymous reader for many helpful suggestions and especially for skillful excisions which sharpened this paper and provided much grist for additional work.

[1] The traditional literature came of age with Pigou's [15] *Economics of Welfare*, which extended Marshall's treatment [11] and Young's [23] constructive review of Pigou's earlier *Wealth and Welfare* before the first World War. It reached its first plateau in the 1920's with the controversy among Clapham [5], Robertson [17], and Pigou [14] over "Those Empty Economic Boxes" and Knight's article [9]. Viner [21] systematized the literature in the early 1930's. Ellis and Fellner [8] and Lerner [10] made contributions during the 1940's. The first satisfactory statement distinguishing between technological and pecuniary economies, Samuelson notwithstanding [18, p. 209], seems to have been made by Baumol [1], although he does not claim originality.

[2] The recent analysis seems to have begun with Meade [12] and has been continued by Scitovsky [19], Buchanan and Stubblebine [3], Davis and Whinston [6], Turvey [20], and Wellisz [22]. It is well summarized up to 1965 by Mishan [13], but continues, in the work of Plott [16], Buchanan [2], Dolbear [7], and others.

addition to discrepancies between the ratios of marginal products to opportunity costs, the firms' average and marginal cost curves shift upwards as the total demand for some resource rises thus generating a rising industry supply price.[3] Technological economies exist when the firms' curves fall under these circumstances. Pecuniary economies (diseconomies) exist when the industry average cost is falling (rising) but the marginal equalities between factor prices which reflect alternative social costs and their marginal products prevail. (2) The size of the welfare loss associated with economies and diseconomies (henceforth referred to simply as economies unless otherwise specified) is very much affected by the particular type of externality present and the structure of the industry in question. The types of externalities considered are pecuniary or technological and the types of industry structures are competition, simple and first degree discriminating monopoly, and simple and first degree discriminating monopsony. (3) It is useful to distinguish economies that are internal to an industry from those that are external to it because some industry structures offset technological externalities when they are internal to a specific industry, but cannot when they are external.

The plan for this paper is as follows: (1) A linear homogeneous production function is described which will provide the basis for theoretical analysis. (2) The analysis relevant to a competitive industry experiencing technological diseconomies is examined in detail so as to provide a solid basis for the treatment of other industry structures and for economies. (3) The advantages and disadvantages of a few of the alternative ways of harmonizing private and social costs are examined. Analysis of specific cases, however, is left to another occasion.

I. *The Production Function*

The following analysis is illustrated by examples based upon a production function homogeneous of degree one, which displays ridge lines in the positive quadrant and which is considered to be relevant to a whole industry whether organized competitively or otherwise. This simple, comparatively well known function is chosen to emphasize the fact that neither technological nor pecuniary economies or diseconomies rest fundamentally upon the nature of the returns to the industry, but rather rest upon violation of the equality of the ratios of marginal products to their factor prices at equilibrium positions.[4]

Figure 1 consists of four panels which

[3] "Industry supply curve" refers to the locus of minimum average cost for a group of one or more firms producing a homogeneous product where the quantity of all factors (which are assumed to be homogeneous) and the numbers of firms are freely variable as industry output expands except that the total supply of some input or inputs may be fixed to the industry. It may be defined alternatively as the summed output of all firms in the industry at each price where the long-run and the short-run marginal cost are both equal to each alternative price. This is often referred to as the long-run supply price (*LRSP*) curve. "Industry marginal cost" (MC_I) is similarly construed, and is marginal to the industry average cost. Under competitive conditions, MC_I is a decision variable for no firm although it denotes marginal social cost (*MSC*) when a technological externality is present.

[4] This point seems to have been made first by Abba P. Lerner in 1943 [10, Ch. 15, 16, 17] but, so far as I know, has not become part of the literature. Lerner did not thus distinguish between technological and pecuniary (dis)economies because his "Rule" is carefully designed to price indivisible and fixed factors according to what might now be termed their shadow marginal products. The present analysis attempts to show that this procedure converts technological into pecuniary diseconomies and thereby supports Lerner's conclusion that misallocation under these circumstances is nothing more than a failure to adhere to the equality of the ratios of factor prices to their respective marginal products. Lerner's handling of technological economies is just as advanced (virtually the 1968 level), if not so satisfactory. Acceptance of Lerner's procedure may have suffered because it directs attention to his proposed social control devices rather than his analysis, but fundamentally it seems simply to have been too neat.

describe the production function, the costs and revenues, and the demand for factors of production by an industry that will alternately be considered to represent each of five industry structures. Panel A illustrates a production function of the postulated type with two inputs, labor (L), and land (T). Since proportionate increases in inputs yield proportionate increases in output, all outputs are readily determined if the output index is taken as 1.0 where isoquant Q_0 crosses ridge line OL. Land represents some input whose supply is fixed and which is useful only in the industry under consideration. Fixity of supply is not essential to the argument but is a useful simplifying assumption relevant to a number of important industries. Usefulness to but one industry, while not essential to the principal conclusion, is not only relevant to important industries but involves significant theoretical and practical distinctions. Removal of this restriction generates pecuniary externalities, which I wish to ignore. Panel B shows the relationship between output and total cost. It also provides a money scale that permits one to show alternative total revenue functions which correspond to alternative prices. The usual geometric relationships are used to find the industry marginal and average money cost curves shown in Panel C. They are drawn on the assumption that the price of land, W_T, is zero, and the price of labor, W_L, is positive at the level illustrated in Panel D. Rising factor costs to the industry are considered later when pecuniary economies are analyzed.

The lines on Panel D reveal the marginal products of both labor and land, calculated in principle for land as a shadow marginal product implied by the marginal rates of substitution at the intersections of the various isoproduct curves with the expansion path TT_1. The input axis of Panel D shows the proportion between the variable factor (labor) and the fixed factor (land) associated with each level of output. If price is a constant at P_0 (see Panel C), these curves also depict the value of the various marginal and average products, and provide a scale upon which money wages and rents can be shown.

No panel shows the position of an individual firm when the industry is viewed as being competitive. The average cost for the industry (AC_I) is identical to the conventional long-run supply curve (referred to henceforth as *LRSP*) traced out by the loci of the average cost associated with the intersections of the long-run marginal cost and short-run marginal cost as the number of firms fluctuates in response to changed industry demand.[5] Whatever the industry structure, this curve is considered to represent minimum cost of production for each output when factor prices are fixed at W_L and W_T (Panel D).

The *LRSP* of a competitive industry as shown by the *LRSP* in Panel C may be said to reflect technological economies from zero output up to Q_B, and technological diseconomies beyond that point, because with constant factor prices the U-shape depends solely upon the production function for the industry. A corresponding function with the same shape reflects pecuniary economies where an optimal rent is charged. The size of the rent must vary as the industry price and optimal output varies. It will be a charge sufficient to make the private factor cost equal to its social cost.

II. *Competitive Equilibrium for a Sub-Industry*

To facilitate a clear distinction between rents and profits, consider first a sub-industry, catching a particular kind of fish from a particular fishing ground. It is assumed to comprise so small a part of the

[5] Point B is marginal to *LRSP* at point C. The marginal cost to *LRSP* at point B is far higher and is shown inexactly as B'.

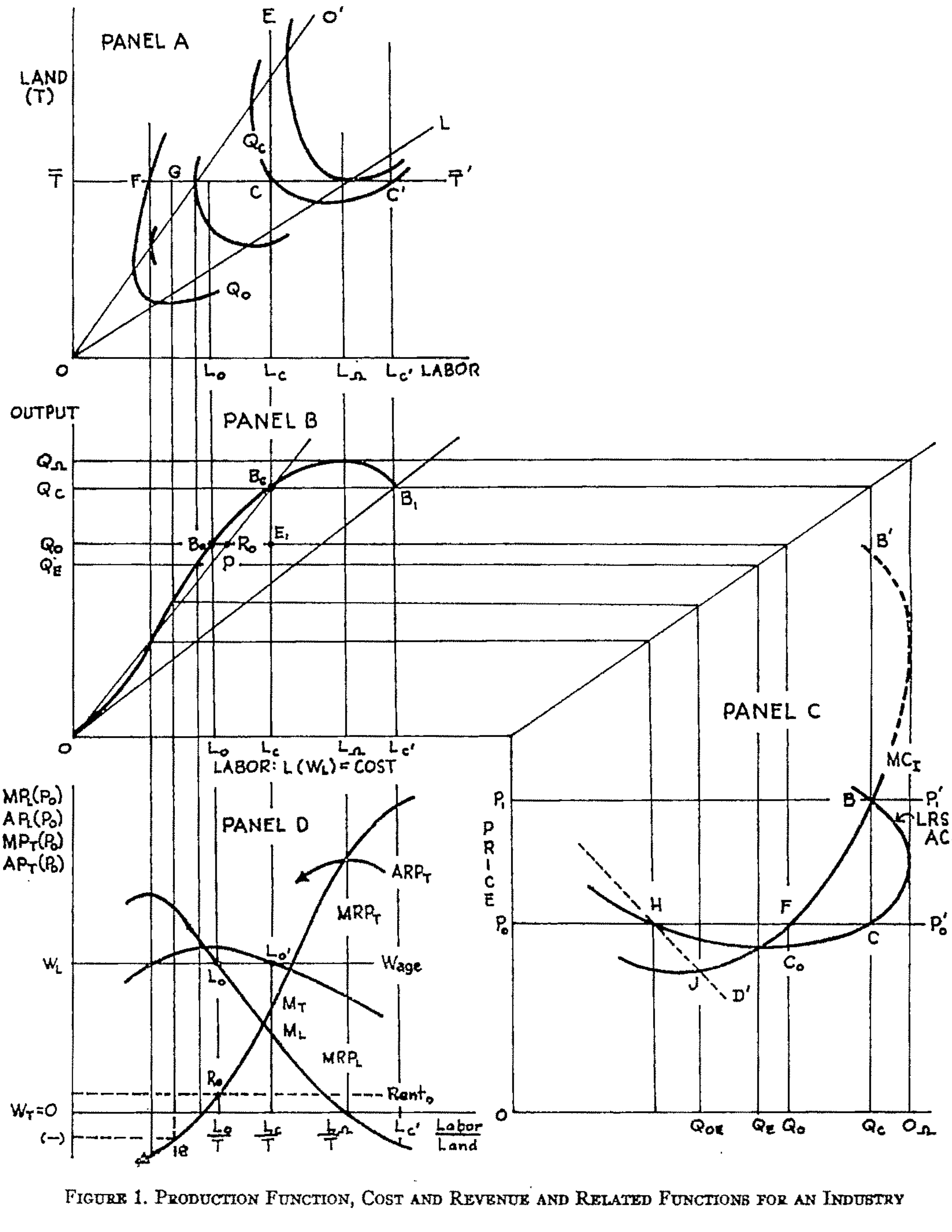


FIGURE 1. PRODUCTION FUNCTION, COST AND REVENUE AND RELATED FUNCTIONS FOR AN INDUSTRY

market for that species that its price is affected overwhelmingly by the demand for fish in general and only imperceptibly by its own output. Two of these narrowly defined industry demand curves are shown in Panel C, namely P_0P_0', and P_1P_1', and in Panel B by the corresponding total revenue curves, OB_c and OB_1. Such a sub-

industry will expand output until profits fall to a level just sufficient to hold the firms in the industry, their summed outputs equalling the quantity demanded. The equilibrium output is in this case the same for both of these demand functions as is shown by Q_c (Panels B and C) and points c and c' on isoquant Q_c in Panel A. Where labor inputs exceed L_Ω (Panels A, B, and D), this competitive industry exhibits a true backward bending supply curve. Such have been observed in certain fisheries and oil fields with multiple owners, and probably exist for other common property resources and publicly supplied goods.[6] This happens because in the absence of a charge for land, the average value product of labor for each firm is equal to the wage rate at equilibrium, although the marginal factor cost to the industry is much higher. In this instance, all costs are labor costs, and the industry expands or contracts until abnormal profits are eliminated. This occurs where price = *LRSP* in the product market (point C in Panel C), although marginal cost to the industry is at point B.

At the higher price, P_1, valuable resources with positive opportunity costs are employed, specifically additional labor in the amount of L_cL_c' (Panels A and B), with no net addition to total production, the marginal product of labor being positive at first but negative beyond L_Ω. Lesser, but still substantial waste occurs if price is P_0, output Q_c, and inputs L_c/T. Attention is henceforth confined to the analysis of this lesser amount of waste.

It is immediately evident from Panel D that, where inputs are L_c/T, the marginal product of labor to the industry point M_L is a bit less than 50 percent of the wage rate, and that the marginal product of land, M_T, is far above the (zero) price at which it enters the firms' cost functions. This relationship is also shown in Panel A by the negative slope of the industry isoquant at point C where isoquant Q_c is not tangential to the isocost line L_cCE. The latter is vertical because the price of land is zero. The inevitable conclusion follows (however paradoxical it might have seemed to Ricardo or Marx). Although labor gets the whole return, social output is much reduced because labor is being wasted for want of price on land. Labor is being wasted because its alternative marginal contribution to other industries exceeds its marginal value product here. Output should, of course, be contracted until labor's marginal productivity to the industry (rather than its average productivity) equals its wage. This occurs at point L_0/T in Panel D and corresponds to an optimal output of Q_0 in Panels B and C. At that point, a positive rent exists since the marginal product of land is R_0 (Panel D). This is equivalent to the difference between the average and marginal products of labor (Panel D), the gap between total revenue and total cost, B_0R_0, shown in Panel B, and the difference between *LRSP* and price times the output, C_0F (P_0F) in Panel C.

Ownership of the land will tend strongly to produce an optimal level of production. In this case, where output of the subindustry does not affect the output price, it does not matter whether ownership of land is concentrated or dispersed.

This treatment seems one-sided since the classical remedy via taxation is ignored. In

[6] An example of publicly supplied goods is streets, especially freeways that are so overloaded at peak hours as to suffer a reduction of traffic flow. If possible, rents should be charged during these periods without further impeding traffic flow. Such a system apparently is technically feasible. I have been told that the Netherland's traffic police use electronic devices which photograph speeding cars on certain freeways. Fines are charged against their owners, the photographed license plate providing the needed information for sending the bill. If such a system exists, it could be expanded to photograph all cars at rush periods; those choosing to use the freeway at that time would be billed for the rent.

the case under discussion, taxation, either of inputs or outputs is, in principle, an effective means for optimizing output. An optimal license type tax on inputs is equivalent to rents. A tax on the output sufficient to reduce the net price to suppliers to C_0 (Panel C) would also result in optimal output. Taxation is widely regarded as superior to rents because the tax revenues supposedly have more desirable income effects than do rents. There are disadvantages in administration, however, for it is also true that the tax rate, whether levied against inputs or outputs, must be optimally adjusted for every change of product price, factor costs, and technology that affects factor proportions. One can believe that the adjustment process via changing rents will be more finely tuned than via tax adjustment.

III. *Welfare Loss Due to Failure to Charge Rents*

The welfare loss resulting from a failure to charge the optimum rent in the case illustrated by Figure 1 can be shown by either of two areas in Panel C, or by a comparison of cost and revenue in Panel B. If price is P_0 and if no rents are charged, equilibrium output is Q_c and both total revenue and total cost for the industry are Q_cB_c (Panel B), or Q_cCP_0O in Panel C.[7] Yet, marginal social cost, which is equal to MC_I in Panel C, is Q_cB, about 60 percent higher than P_0. If maximum rents are charged, output drops to Q_0 and unit costs to Q_0C_0 (Panel C) where the social optimum is indicated by the intersection of MC_I with P_0 at point F. Total costs (excluding rents) and revenues are equal to Q_0B_0 and Q_0R_0 respectively on Panel B.

Comparison with the competitive equilibrium shows that consumers lose satisfaction for which they would have paid R_0E_1 (Panel B), while resources valued at B_0E_1 are saved and presumably used to produce an equivalent value of goods and services in other industries. The net gain, which is the largest attainable, is B_0R_0: precisely equal to the maximum rent.

The same result is illustrated, perhaps more persuasively, in Panel C where the rent and the welfare gain appear as different areas. The maximum rent is shown by the rectangle with sides P_0F and FC_0. The welfare gain resulting from the imposition of the rent is shown by the approximately triangular area FCB. The fact that these two areas must be equal appears obvious from Panel B, but it holds only when price is equal to marginal revenue for the industry. Otherwise the rectangular area includes a combined maximum of rent and profit, which occurs when output is restricted below the optimum level and the demand curve is sloping. The profit plus rent rectangle is augmented, but the welfare gain is reduced. This situation is analyzed below.

IV. *"Technological" Diseconomies:*

Competition, Normal Industry Demand

Now that the flat sub-industry demand curve has enabled us to define the marginal social cost function and to relate it to rents, it is time to admit that even sub-industries typically confront demand curves that slope downward and toward the right as illustrated in Figure 2, which is otherwise the equivalent of Panel C. The relationships among physical inputs and outputs, and the prices of the factors remain the same as in Figure 1, and the now sloping demand curve is drawn to leave the no rent competitive price P_0 and output Q_c unchanged. The money demand for factors as a whole would (if shown) be rotated clockwise around their former equilibrium positions.

The principal modification of the analysis is the reduction in the size of the welfare gain incident to the imposition of an opti-

[7] It should be kept in mind that each firm is in zero pure profit equilibrium with $MC=AC=P$ when the industry operates at Q_c, or indeed at any other output where $LRSP=P$.

mal rent to the shaded area CBP_r, and an increase of optimal rent to P_rE per unit of output. In this case, optimal rents are approximately doubled while the potential welfare gain is lessened. Each of these modifications follows from the fact that a smaller adjustment of output below the competitive level suffices to achieve the optimal result, since price rises as output is reduced so that optimum output and price occur at Q_{0M} and P_r.

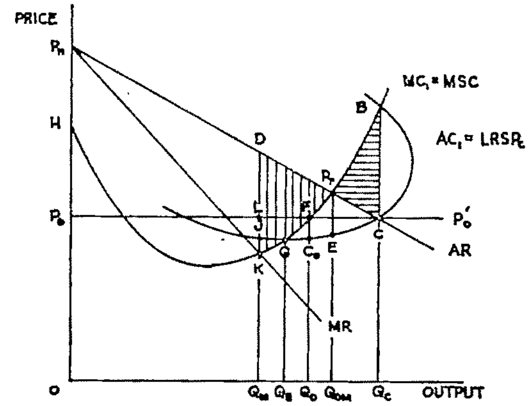


FIGURE 2. OPTIMA: TECHNOLOGICAL DISECONOMIES

Simple and Discriminating Monopsony

The same situation illustrated by Figure 2 can also be viewed as a monopsony. Suppose that the freely competing fishermen must sell to a single cannery (perhaps a local government monopoly). The cannery may be completely devoid of monopoly power, selling its products in distant markets in competition with hundreds of rivals. In this case the appropriate demand curve for the fish is the horizontal line, P_0P_0'. A maximizing simple monopsonist will buy quantity Q_0 at price C_0 which it resells at price OP_0 plus its value added per unit. His profit is equal to FC_0 per unit of output, exactly the maximum rent under competition. Output is induced to the optimal level because his profits are maximized where MC_I (the marginal social cost in this case) equals price. The price paid to the fishermen, C_0, is on the *LRSP* for the industry, so only the correct number of firms and levels of output are induced. If the demand curve is P_nC, the competitive optimum (with rents) is attained at output Q_{0M}, and fishermen's price E.

A discriminating monopsonist, on the other hand, will overproduce. This is illustrated by the extreme case of the perfectly discriminating monopsonist who drives all or nothing bargains with each fisherman, or pays a minimum price for each unit purchased thus making the discriminating monopsonist's marginal cost curve equal to his average cost curve which is the *LRSP* for the industry as shown in Figures 1 and 2. A discriminating monopolist maximizes profit when output is carried to the competitive no rent level, Q_c. Welfare loss (FCB) is the same as under competition; profits exceed optimal rent slightly more than FC_0C.[8]

Simple Monopoly[9]

Consider again the effect of the sloping industry demand curve P_nC. Should the

[8] The appropriate profit calculation is somewhat troublesome because external economies are shown up to Q_E and external diseconomies thereafter. In a later section, we find the direction of adjustments reversed for external economies. In the present instance, it seems most reasonable to assume that the discriminating monopsonist will use his power to hold purchase prices to the minimum average cost, G, for supplies up to quantity Q_E. The addition to profit (in addition to the rent equivalent) resulting from perfect discrimination is that shown by FC_0C, plus the difference between buying prices G and C_0 for every unit up to Q_E plus a diminishing sliver of profit between Q_E and Q_0.

[9] The nature of costs under monopoly and monopsony are often insufficiently defined. If long-run monopoly and competitive equilibria are to be contrasted on the assumption of long-lived barriers to entry, then full adjustment to alternative output levels must be permitted. This includes alteration of number of establishments, centralized and decentralized control systems, and the like. When this is done, the average cost curve of the monopolist comes from the same family as the *LRSP* of the competitive industry. The U-shapes which are universally drawn for monopolists are questionable generalizations.

It seems much more likely that they are like the *LRSP* under competition plus an additional cost per unit to cover the additional costs of coordination and preservation of the barriers to entry. Yet, some economies may also be forthcoming. Here the *LRSP* under

firms in this industry form a perfectly efficient cartel such that the control over output becomes absolute and the industry cost function is not raised above that for competition, the cartel would, if a simple monopolist, maximize its profit at Q_M. Price is then $Q_M D$. Because it has control over entry, it will not overuse the resource (which we have assumed thus far to be useful only in this industry), and it will consider the effect of its output on price. Thus marginal revenue will be made equal to marginal social cost.

All of the factors (the price of which we continue to hold constant) including the natural resources are underused by monopolists. Compared to the competitive result when an optimal rent is paid, value of output to the consumers is reduced by the area under DP_r while costs are reduced only by the area under KP_r. There is a welfare loss of the shaded area DP_rK to be set against the welfare gain of P_rBC, the other shaded area, as compared to competitive equilibrium when no rent is paid. Simple monopoly may be either better or worse than competition if no rents are charged depending upon the relative elasticities of the industry demand and of the long-run average cost curve of the monopolist, assuming the latter to be identical to the long-run supply price for the industry when organized competitively.

In this example, rent per unit of output falls to JL, but profit per unit of output is LD. In either case, the gains, if any, to the consumer are small compared to the gains in profits plus rents to the cartelizers, when both are measured in money. One must imagine brutish consumers and epicurean businessmen to be willing to urge the merits of increased efficiency by monopolization even in those cases where there is a clear welfare gain. A government monopoly might find approbation if the industry is thought to be essential and if the profits plus rents were utilized to accelerate economic growth, but it is not a welfare optimum.

A sophisticated combination of subsidies that effect marginal costs and induce larger output and a license type tax to capture pure profits or rents can, in principle, produce an optimum result under simple monopoly. The relevant curves must be accurately known, and prompt adjustments made when they shift if an optimum is to be attained.

Perfectly Discriminating Monopoly

A perfectly discriminating monopolist, whose long-run average cost is identical to the competitive long-run supply curve, would carry output to the socially optimal level (Q_{0M} in Figure 2) provided that variable inputs' wages are determined independently of industry output and that fixed inputs are specific to the industry, as is the case here.

Implicit rent is restored to the maximum level (EP_r per unit of output) and profit plus rent is maximized (P_nP_rKH). The reason for the good allocation result is that all social costs are internalized by the decision making unit when such an industry is monopolized, and the demand curve is also the industry marginal revenue curve. This is true with perfect discrimination since the firm need not cut the price on any unit of sales in order to extend its sales further by means of lower prices to otherwise excluded buyers. The transfer to the monopolist is enormous and may be more objectionable to many social critics than the misallocation that is overcome, but this is a different problem.

V. *Pecuniary Diseconomies*

Virtually all of the conclusions which relate differing industry structure to opti-

competition and the LAC under monopoly are treated as identical, and analysis is confined to the effect of structure on the conditions for equilibrium and their effect on optimal resource allocation. A U-shaped $LRSP$ is used only for illustrative purposes and to reduce the number of figures, since that shape contains all three types of returns.

mal allocation in the presence of technological diseconomies must be reversed when pecuniary diseconomies cause cost changes as an industry expands or contracts. Optimal allocation of resources occurs under competition and perfectly discriminating monopsony; restricted production and welfare loss is a consequence of simple monopoly, perfectly discriminating monopoly, and of simple monopsony.

The textbook distinction between technological and pecuniary diseconomies is that the latter is accounted for by rising factor costs per unit of output, not the nature of the production function. The preceding analysis has shown this distinction to be but a special case of a more fundamental difference. Technological diseconomies occur when the cost of output is rising and the cost of some input to the firms is less than its opportunity costs to its industry without reference to the shape of the factor supply curves. Hence any rise in output cost is pecuniary if the marginal equalities are maintained between the prices of factors which reflect opportunity costs and their respective marginal products. Whenever this is true, the *LRSP* is the curve of marginal social costs (*MSC*), and optimal output is found where *LRSP* $=P$. This is true because any increase in the marginal cost of output per unit reflects either (1) a rise in the opportunity cost of the factors due to expansion of this industry which thereby raises their opportunity costs in each of their uses, or (2) because of lower marginal productivity, or both.

This may be illustrated with Figures 1 and 2 when optimum rents are charged. The *LRSP* for a competitive industry becomes identical to the *MSC* (which, it will be recalled, coincides with the former MC_I). These are shown in Figure 3 as $LRSP_{pec}$ and MC_{tech}. The *MSC* curve, of course, does not shift.

A new marginal cost to the industry arises which is marginal to the new *LRSP*.

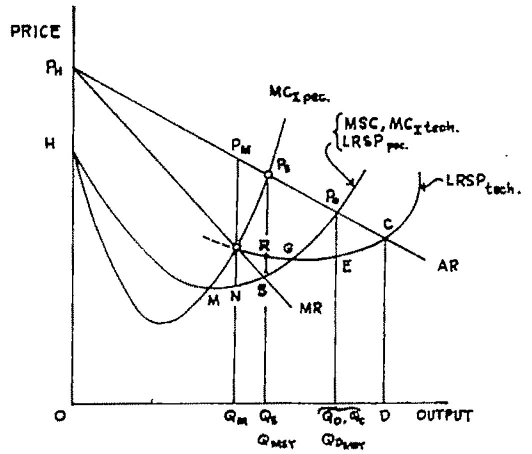


FIGURE 3. OPTIMA: TECHNOLOGICAL AND PECUNIARY DISECONOMIES

It is denoted MC_{pec} in Figure 3, and is relevant to pecuniary externalities.

The welfare effects relevant to the five alternative market structures are readily compared using the standard definitions of equilibrium. Optimal output occurs at Q_0 where $MSC=$ Demand. This is also the competitive equilibrium because $LRSP_{pec} \equiv MSC$. Simple monopoly operates at Q_M where $MR=MC_{pec}$ and produces a welfare loss of $P_M P_0 N$. Unlike the situation where the diseconomies are technological and monopolistic restriction tends to offset competitive overexpansion, this is an unmitigated, even a multiplied, loss because the firm's marginal cost lies above marginal social cost and its marginal revenue lies below the demand curve in the relevant ranges. A perfectly discriminating monopolist maximizes where $MC_{pec}=$ Price. A welfare loss of $P_s P_0 S$ results since MC_{pec} lies above *MSC*.

Viewed as a simple monopsony, the supply curve to a single buyer in this competitive market is $LRSP_{pec}=MSC$. The buyer's marginal cost is, therefore, MC_{pec} which he will set equal to his demand curve, and buy Q_{msy} at price S.[10] Again the

[10] The reader may note an anomaly here. $LRSP_{pec}$ is taken as the supply price even to the left of point *G* where the optimum rent is zero. This implies an opti-

welfare loss is $P_s P_0 S$. A perfectly discriminating monopsonist will find his marginal factor cost to be equal to the $LRSP_{pec}$. Profit maximization occurs at Q_0 which is the same as the competitive output and is a welfare optimum according to the usual definitions. Profits, however, are not normal, being equal to $P_H P_0 MH$.

All of the noncompetitive structures garner profits, but the relationship between the size of the profits and the size of the welfare loss is highly irregular.

VI. *External Economies*

Figure 1 illustrates too, the welfare results of alternative structures for external technological and pecuniary economies. The dashed line in Panel C is interpreted as the industry demand curve. The cost curves of Panel C up to Q_B, and the corresponding portions of the associated curves in the other panels in Figure 1, relate the alternative equilibria to the basic technological and factor price constraints where external economies accompany the expansion of the industry. Market demand is such that with the most efficient organization of production, competitive equilibrium must occur at point H (with zero price of land) at which point $LRSP$ is still falling. Tracing this equilibrium back to Panel A, we find point F which is outside of the economic sector because the slope of the isoquant denotes a negative marginal product for land. Tracing this combination to Panel D, we discover not only that the marginal product of land is well below its (zero) price to the firms in the industry, but that the marginal product of labor is correspondingly far above its wage rate. Labor is being wasted by insufficiently intensive use. Yet profits are only normal.

Point J (Panel C) represents optimal allocation where demand $= MSC$. If its associated output, Q_{0t}, is traced back to Panel A, we find a smaller but still negative marginal product of land implied by the interpolated slope of the isoquant at point G. The corresponding points in Panel D seem to suggest a continued failure to bring the marginal cost of labor into equality with its wage rate, but this is an illusion. The price of the product is about one-fourth lower at J, as compared to H, so that the value of the marginal product in Panel D should be altered proportionately. This produces the desired equality.

The value of the marginal product of land requires a similar adjustment, but it must remain negative at the optimal output and combination of inputs. Thus, the optimal marginal product of land must be negative, and a competitive industry can produce at the optimal rate only if a negative rent (a subsidy) of optimal size be paid for the use of the unpriced resource.

The need for a subsidy to competitive industries to induce optimal levels of operation in the presence of technological economies has long been known. The very similar argument favoring subsidization of regulated monopolies when economies of scale are internal to the firm is based on the same analytical conclusion, that profitable operation is impossible at the socially optimal rate of output.

The equilibrium positions for other industry structures where external technological or pecuniary economies exist are illustrated in Figure 4 which reproduces Panel C and also includes the cost curves relevant to pecuniary economies. Subscripts $_t$ and $_p$ are used to designate points relevant, respectively, to technological and pecuniary externalities. The analysis of all four conditions, technological and pecuniary diseconomies and economies, is summarized in Table 1.

It is evident that a given structure yields the same general welfare effect whether

mum negative rent where $LRSP_{pec} < LRSP_{tech}$. This is indeed the correct implication. In this range, technological economies exist which involve negative shadow marginal products for the unpriced fixed factor. This case is discussed in the next section.

TABLE 1—EQUILIBRIUM CONDITIONS AND WELFARE LOSS DUE TO EXTERNAL DISECONOMIES AND ECONOMIES, BY TYPE OF INDUSTRY STRUCTURE

Structure	Equilibrium Condition (all cases)	Diseconomies (See Fig. 3)		Economies (See Fig. 4)	
		Price Quantity	Welfare Loss	Price Quantity	Welfare Loss
Competition					
Technological	$LRSP_t=P$	C, D	P_0CE	P_e, Q_e	P_eP_0E
Pecuniary	$LRSP_p=P$	P_0, Q_0	optimum	P_e, Q_0	optimum
Simple Monopoly					
Technological	$MC_t=MR$	P_s, Q_s	P_sP_0S	P_{MT}, Q_{MT}	$P_{MT}P_0K$
Pecuniary	$MC_p=MR$	P_m, Q_m	P_MP_0N	P_{MP}, Q_{MP}	$P_{MP}P_0K'$
Discriminating Monopoly					
Technological	$MC_t=P$	P_0, Q_0	optimum	P_0, Q_0	optimum
Pecuniary	$MC_p=P$	P_s, Q_s	P_sP_0S*	P_s, Q_s	P_sP_0E'
Simple Monopsony (where demand is seen as marginal value product and *LRSP* as the supply function)					
Technological†	$MC_t=MFC=P$	E, Q_0	optimum	W, Q_0	optimum
Pecuniary	$MC_p=MFC=P$	S, Q_s	P_sP_0S*	E', Q_s	P_sP_0E'
Discriminating Monopsony					
Technological	$LRSP_t=P$	C, D	P_0CE	same as simple monopsony**	
Pecuniary	$LRSP_p=P$	P_0, Q_0	optimum		

† Note that the wage lies on the supply curve *LRSP* even when the supply curve lies above the demand curve.

* This is equal to the loss for a simple monopolist under technological economies only by coincidence.

** Discriminating monopsony yields lower profits than simple monopsony when the supply curve is downward sloping because discrimination raises the average price of the factor. Hence a firm with power to discriminate will not utilize it.

such external effects are diseconomies or economies, and that a structure that optimizes when economies are pecuniary will fail to do so when they are technological and vice versa.

VII. *Externalities External to the Industry*

One important limitation has been retained thus far, the industry concept and the implication that externalities exist only among firms producing a homogeneous product. When this limitation is removed, choice among means for correcting for externalities is reduced in principle and virtually eliminated as a practical matter. One can argue persuasively that optimization by simple monopsony or by organizing perfect monopolistic discrimination is entirely impractical in any case even if income effects should be ignored. If the economies or diseconomies are external not only to the firms in an industry, but also to any given industry, i.e., among firms that are not producing homogeneous products, the structural options are lost

even in principle. An optimizing policy must attempt to achieve competitive equilibrium with inputs priced according to their alternative costs. This is true because it is virtually certain that the factor proportions of different users will differ so that the improperly priced factor will not be properly allocated by discrimination or regulation among competing users. The foregoing analysis assumes that the improperly priced factor has but one use and will be given an optimal implied value by a monopolist who treats it as a residual. This implied price is the rent. If the improperly priced factor is used in two or more industries, its allocation among them needs to be such that its implicit marginal value to each industry is the same.[11]

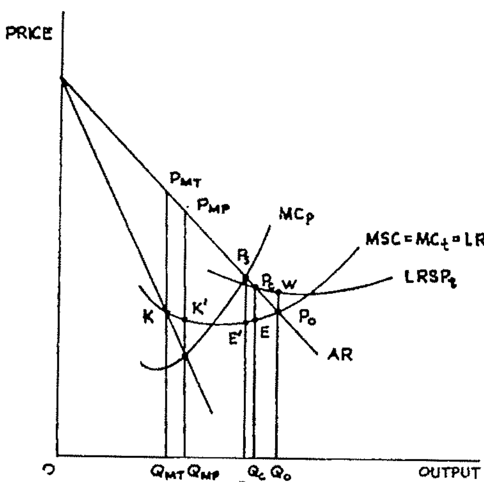


FIGURE 4. OPTIMA: TECHNOLOGICAL AND PECUNIARY ECONOMIES

VIII. *Are Any Boxes Empty?*

It is now possible to list the various types of external economies. They fall conveniently into eight categories. Each is listed below with an illustrative example. This classification should not obscure the more basic fact that any technological economy or diseconomy can be converted into a pecuniary one by appropriate pricing of inputs. The listing reflects the author's judgment of the typical pricing practices in the United States. These should be examined more carefully after alternative methods of adjusting for technological economies and diseconomies have been surveyed and refined.

A. Externalities which are internal to a specific industry:

1. Technological diseconomy: common property resources; fisheries, and oil formations used by competing firms.
2. Technological economy: public utilities; especially electric power, water supply, sewage disposal where a single price covers all costs.
3. Pecuniary diseconomy: the most common case, where expansion encounters rising factor prices and/or declining marginal productivity.
4. Pecuniary economy: increased specialization via purchased inputs or subcontracting with expansion of industry output; clothing, aircraft.

B. Externalities which are external to a particular industry:

5. Technological diseconomy: most public services which are supplied without a charge to the user according to use; city streets, air.
6. Technological economy: basic research, exploration, law and order, provision of information service.
7. Pecuniary diseconomy: there are many of these; high cost of living, and wages in the larger cities as city size increases.
8. Pecuniary economies: general specialization.

The same industry will often appear in both type A and type B. Thus, the internal technological economy for electric power

[11] This point is also valid within an industry when the various firms utilize different processes and have different implied values for the input fixed to the industry.

production (#2) can also be cited for external pecuniary economy (#8).

IX. *Summary*

All external effects are found in principle to be reducible by correct input pricing to economies or diseconomies. This holds for separable and nonseparable externalities in production and for externalities viewed either in terms of specific firms or among industries. Different industry structures produce optimal allocation either for technological or for pecuniary externalities if the externalities are internal to an industry and the firms therein use the same factor proportions. When these conditions do not hold, only competitive structures and pecuniary externalities suffice to produce an optimum in the absence of very complicated combinations of taxes and subsidies which require a wealth of accurate detailed data utilized with consummate insight and speed.

References

1. W. J. Baumol, *Welfare Economics and the Theory of the State*, Cambridge 1952.
2. J. M. Buchanan, "Joint Supply, Externality and Optimality," *Economica*, Nov. 1966, *33*, 404–15.
3. ———, and W. C. Stubblebine, "Externality," *Economica*, Nov. 1962, *29*, 371–84.
4. R. Coase, "The Problem of Social Cost," *J. Law Econ.*, Oct. 1961, *3*, 1–44.
5. J. H. Clapham, "Of Empty Economic Boxes," *Econ. J.*, Sept. 1922, *32*, 305–14; reprinted in K. E. Boulding and G. J. Stigler, *AEA Readings in Economic Theory*, Vol. 6, Homewood 1952; and Clapham, "Rejoinder," *Econ. J.* Dec. 1922, *32*, 560–63.
6. O. A. Davis and A. Whinston, "Externalities, Welfare and the Theory of Games," *J. Polit. Econ.*, June 1962, *70*, 241–62.
7. F. T. Dolbear, Jr., "On the Theory of Optimal Externality," *Amer. Econ. Rev.*, Mar. 1967, *57*, 90–103.
8. H. Ellis and W. Fellner, "External Economies and Diseconomies," *Amer. Econ. Rev.*, Sept. 1943, *33*, 493–511; reprinted in *Readings* [5].
9. F. H. Knight, "Fallacies in the Interpretation of Social Cost," *Quart. J. Econ.* Aug. 1924, *38*, 582–606; reprinted in *Readings* [5].
10. A. P. Lerner, *The Economics of Control*, New York 1943.
11. A. Marshall, *Principles of Economics*, 8th ed. London 1920.
12. J. E. Meade, "External Economies and Diseconomies in a Competitive Situation," *Econ. J.*, Mar. 1952, *67*, 54–67.
13. E. J. Mishan, "Reflections on Recent Developments in the Concept of External Effects," *Can. J. Econ.*, Feb. 1965, *31*, 1–34.
14. A. C. Pigou, "Empty Economic Boxes: A Reply," *Econ. J.*, Dec. 1922, *32*, 458–65; reprinted in *Readings* [5].
15. ———, *The Economics of Welfare*, 4th ed., London 1952.
16. C. R. Plott, "Externalities and Corrective Taxes," *Economica*, Feb. 1966, *33*, 84–87.
17. D. H. Robertson, "Those Empty Boxes," *Econ. J.*, March 1924, *24*, 16–30; reprinted in *Readings* [5].
18. P. A. Samuelson, *Foundations of Economic Analysis*, Atheneum ed. New York 1965.
19. T. Scitovsky, "Two Concepts of External Economies," *J. Polit. Econ.*, Apr. 1954, *57*, 143–51.
20. Ralph Turvey, "On Divergencies Between Social Cost and Private Cost," *Economica*, Aug. 1963, *30*, 309–13.
21. J. Viner, "Cost Curves and Supply Curves," *Z. Nationalökon.*, 1931, *3*, 23–46; reprinted in the *Readings* [5].
22. S. Wellisz, "On External Diseconomies and the Government Assisted Invisible Hand," *Economica*, Nov. 1964, *31*, 345–62.
23. A. Young, "Pigou's Wealth and Welfare Review," *Quart. J. Econ.*, Aug. 1913, *27* 672–86.

Dependency Rates and Savings Rates

By NATHANIEL H. LEFF*

This paper is concerned with a theory of aggregate savings rates which offers a partial explanation of the well-known fact that the savings-income ratio (S/Y) is generally higher in the developed countries than in the underdeveloped countries.[1] Most discussions of the determinants of international savings rates have accorded an important role to the level of per capita income. This view is supported by the observation that S/Y is usually higher in the rich countries than in the poor, and has received some support from the statistical studies of international savings rates.[2] These studies, however, raise an important question. The cross section results suggest that with rising income levels in individual countries, aggregate savings rates should rise as well.[3] This has not been borne out, however, by the time-series data for most countries. Per capita income has been growing at annual rates of 1–2.5 percent in most underdeveloped countries during the past 15 years; and many of the social and economic changes, such as expansion of the capitalist sector, which are expected to raise aggregate savings ratios have in fact occurred. Nevertheless, savings rates have generally not shown an upward trend [34, pp. 14–16]. A number of alternative hypotheses can be mobilized to explain this phenomenon; for example, the possibility of an international demonstration effect influencing consumption.[4] This paper considers one particular hypothesis, and presents statistical results to test it. We then go on to consider some of the analytical and policy implications.

I. *Demographic Conditions and Savings*

The cross section and time-series savings functions may be inconsistent, because the relation between S/Y and per capita income (Y/N) is incorrectly specified. The correlation between S/Y and Y/N may, at least in part, reflect a connection between S/Y and another variable which did *not* change as levels of per capita income have risen.[5] Crude birth rates are an ob-

* The author is an assistant professor at the Graduate Business School of Columbia University. This paper was prepared as part of the 1967 Summer Research Project of the Office of Program and Policy Coordination of the U.S. Agency for International Development. The paper does not necessarily reflect the views of *A.I.D.* I am grateful to Thad Alton, Hollis Chenery, Raymond Goldsmith, H. S. Houthakker, Harry G. Johnson, Edwin Kuh, Harvey Leibenstein, H. Lubin, Laura Randall, Alan Strout, and Maurice Wilkinson for valuable comments and suggestions on this project. I bear sole responsibility for any deficiencies.

[1] For the 47 underdeveloped countries considered in the present study, S/Y averaged 14.2 percent in 1964, with a standard deviation of 4.5. For the 20 Western developed countries, the figures were 25.2 percent and 4.8, respectively. This was computed from data discussed below.

[2] See, e.g., Simon Kuznets [16]; Charles Wolf, Jr. [30, p. 19]; United Nations [34, pp. 13–18]; Luis Landau [17]. H. S. Houthakker [15, pp. 220–21] also showed some evidence for slight curvilinearity between income and gross domestic savings.

[3] The results of Alan Strout [28, Tables 3–4] and Hollis B. Chenery and Peter Eckstein [3], also indicated marginal savings rates greater than the average. Modigliani and Brumberg [21], [20], and Milton Friedman [12] have of course maintained on theoretical grounds that the long-run savings/income ratio should be constant. This is discussed below.

[4] This problem is similar to that raised in the 1940's in the United States and discussed by Duesenberry [10], Modigliani and Brumberg [21], and Friedman [12]. Perhaps because of its focus on the (private) consumption function, however, that discussion has tended to concentrate on household savings rather than on the aggregate savings ratio. As discussed below, the consumption function theories cited can be integrated without difficulty with the hypothesis discussed in this paper.

[5] For example, Wolf [30] and Landau [17] report coefficients of multiple determination of .2 to .3 on equations for S/Y and Y/N, as compared with $\bar{R}^2$'s of .8 to .9 for the equations in which per capita savings is

vious candidate for this "true" variable. Ranging from 34–45 per thousand population in most underdeveloped countries as compared with 17–23 for most of the advanced countries, they are highly correlated with per capita income levels.[6] At the same time, birth rates in the underdeveloped countries have generally not declined in the postwar period [31].

The idea that demographic factors should be included in the analysis of savings rates is also indicated on analytical grounds. A number of economists have long suggested that birth rates should be inversely related with a country's savings potential.[7] This suggestion is also supported by casual empiricism. A number of countries, e.g., Japan, Luxembourg, Norway [15, p. 217], and the communist countries of Eastern Europe, which are "exceptions" to a per capita income-savings ratio model, fit nicely into an approach which includes demographic conditions. Their birth rates are all low, relative to their income levels [22]. And although demographic conditions have been suggested by some writers as important determinants of savings rates, this idea has not been in the mainstream of the discussion. Indeed, some studies and textbooks barely mention it. One reason is that the hypothesis has not been validated by extensive empirical testing. Consequently, it should be worthwhile to see if inclusion of demographic factors does increase our capacity to explain international savings behavior.

II. *The Hypothesis*

The notion that high birth rates constrain saving has two components. First, demographic theory indicates that a prolonged high birth rate will affect a population's age composition [5], placing a relatively large percentage of the population in the younger age brackets. This relation is, in fact, borne out in recent demographic experience.[8]

The logic of an inverse relation between dependency ratios and savings rates, in turn, goes as follows. Children constitute a heavy charge for expenditure which, in the standard national income accounting framework, is put under the heading of consumption. Because they contribute to consumption but not to production, a high ratio of dependents to the working age population might be expected to impose a constraint on a society's potential for savings.[9]

These can be thought of as coming from three sources: private household savings, the public sector's surplus on current account, and corporate savings. The hypothesized link between high dependency ratios and low household savings is, as just described, direct.[10] A connection can also be suggested with government revenues, and perhaps surplus if we assume that taxation is levied out of discretionary

regressed against per capita income. In addition, Maurice Wilkinson has pointed out that the high $\bar{R}^2$s obtained in regressions for S/N may occur because of a statistical artifact. Since $S/N = S/Y \cdot Y/N$, the high correlations obtained in regressions of S/N on Y/N may simply reflect the presence of Y/N on both sides of the equation.

[6] The simple correlation between crude birth rates and per capita income for 74 countries is −.71. This was computed from data discussed below.

[7] See, among others: Joseph J. Spengler [26]; Hans W. Singer [25, pp. 80–81]; Milton Friedman [12, p. 123].

[8] Using data for 74 countries [31], a regression of the child dependency ratio (the proportion of the total population aged 14 or less) against the birth rate 14 years earlier (or the rate for the nearest year available to it) showed a correlation of .89 in the 1960's.

[9] The burden of this fixed consumption expense on their families' discretionary expenditure can be considered as being in the same spirit (but the very opposite empirical reality) as the technically dictated investment and savings decisions which Albert Hirschman has emphasized [13, p. 41 ff.].

[10] The hypothesis is not necessarily that larger families save less than smaller families. In some countries, family size is positively related with income; and high income families may indeed be relatively high savers. Standardized for family size, however, their savings might be even higher. Eizenga's study [11] found such a result for personal saving in the United States.

expenditure, which in turn is defined as the surplus above a (conventional) minimum per capita standard of living. Hence, taxation revenues would be lower because the consumption burden imposed by dependents reduces this surplus. There is no obvious reason why the dependency ratio should affect corporate savings; but we should recognize that saving in the corporate, household, and public sectors may be substitutes.[11] Moreover, in countries where corporate ownership is not widespread, the boundary between corporate and family income is not hard and fast, so that the household consumption constraint imposed by dependents would also affect corporate savings. Similarly, in many countries, public sector corporations play an important economic role, particularly in activities which are capital intensive and of large minimum scale. These large investments, which in other countries would be financed from corporate saving, may show up under the institutional heading of a public sector surplus or borrowing from household savings.

The hypothesis of a connection between dependency ratios and savings has in fact been supported by econometric studies which included demographic factors: W. Eizenga's work on personal saving in the United States [11]; and a broader study, Franco Modigliani's pioneering analysis of international "private" (personal plus corporate) savings rates [20, p. 77 ff.].[12] The present work extends the scope of the inquiry to gross domestic savings rates (i.e., including public sector savings). We will also consider a larger number of countries; seventy-four, as compared with the twenty-four and thirty-six in Modigliani's samples.

III. *Statistical Estimation*

The hypothesis outlined in the previous section suggests rewriting the standard national income identity to include separately the consumption of the working age population, C_{Wa}, and the expenditure on the dependents in the population, D. With the *ex post* equality of savings and investment, we have:

$$(1) \qquad Y = C_{Wa} + D + S + X - M$$

Dividing through by Y and rearranging terms, it would follow that S/Y will be inversely related to the share of dependency expenditure in national income. As noted below, the hypothesis can be elaborated theoretically along different lines. It seemed worthwhile as a first step to ascertain whether it has any empirical relevance at all. The model estimated is therefore fairly simple, and seeks to test the proposition that a country's aggregate savings rate is lower, *ceteris paribus*, to the extent that it has more dependents in its population.

The technique adopted was multivariate regression analysis. The dependent variables employed in the different equations are S/Y, the aggregate domestic savings ratio, expressed in percentage points; and S/N, the level of per capita savings, expressed in 1964 dollars. The independent variables used are:

First, the level of per capita income, Y/N, expressed in *U.S.* dollars. The observations for most countries are for 1964.

[11] Modigliani [20, pp. 98–100] adopted a similar line or reasoning with regard to substitutability between corporate and personal saving, and brought strong statistical evidence in support of this hypothesis. I am grateful to him for making his paper available.

[12] Modigliani [20, p. 77 ff.] approached the question of population growth and savings from a somewhat different slant. In accordance with the life cycle hypothesis of savings, he hypothesized that (private) savings should be a positive function of population growth. This would follow as the size of successive age cohorts entering the economically active population, and saving for their retirement, increases over time. However, having failed to obtain statistical support for this hypothesis, he went on to specify tests based on the dependency burden and found results similar to those reported below.

Details on the sources for the data used are presented in the Appendix.

A second independent variable is the rate of growth of income. This is suggested in earlier work by Houthakker [15, p. 222], Modigliani [20, p, 55 ff.], Swamy [29], and Mizoguchi [19] (the latter two on personal saving). The original Modigliani-Brumberg hypothesis suggested using the rate of growth of aggregate income [21]; but preliminary results indicated that the rate of increase of per capita income was much more significant. For this variable, g, the average annual growth rate in the preceding five years, expressed in percentage points, was used.[13]

A third independent variable is D_1 the percentage of the population aged 14 or less. It is a rough index of the dependency burden imposed by children.[14] This has a range of 40–46 percent for most underdeveloped countries, and 24–31 percent for the advanced countries. In accordance with theoretical suggestions, supported by Modigliani's empirical results, that the retired, older population also constitutes a dependency burden by being claimants on consumption without contributing currently to output, an additional variable, D_2, was introduced.[15] This was defined as the percentage of the population aged 65 or older. This variable ranges from 3 to 5 percent in most underdeveloped countries, and 8 to 13 percent in the advanced countries. Finally, the total dependency ratio, D_3, the sum of D_1 and D_2 was used. Demographic, savings, and income data were available for 74 countries: 47 underdeveloped, 20 developed Western countries, and 7 Eastern European communist countries.

Regression Results

Equations (2) and (3) show the results of log-linear regression equations fitted by least squares to the data for 74 countries. Student t-values are presented in parentheses below each coefficient. These are followed underneath by the *beta* coefficients, the statistic which shows the relative importance of the different independent variables in accounting for the variance in the dependent variable.

$$
\begin{aligned}
(2)\quad \ln S/Y = \ & \underset{\substack{(5.7289)\\ -}}{7.3439} + \underset{\substack{(2.8776)\\ .3644}}{.1596 \ln Y/N} \\
& + \underset{\substack{(3.2792)\\ .2570}}{.0254 \ln g} - \underset{\substack{(4.6406)\\ .7977}}{1.3520 \ln D_1} \\
& - \underset{\substack{(2.5623)\\ .5107}}{.3990 \ln D_2}
\end{aligned}
$$

$$\bar{R}^2 = .5697 \qquad F = 25.1604$$

[13] An identification problem may clearly exist in deciding whether the rate of saving is being influenced by the rate of growth or whether the causation goes from g's dependence on the rate of investment and thus on saving. Modigliani [20, pp. 68–75] considered a somewhat similar problem (different in that he dealt with private rather than aggregate domestic saving), and concluded from statistical tests that the relation of savings to growth can be identified. In any case, as we will see below, g turns out not to be very important quantitatively as an independent variable in most of our equations.

[14] The volume of dependency expenditure will depend on the number of dependents and on the standard of living provided for them. Most dependency expenditure, particularly for children, is provided through the institution of the family. As Duesenberry, and Okun [9, pp. 233–34, 236] have pointed out, this circumstance puts bounds on the dependents' standard of living, relating it to the consumption level of their families. In the absence of extensive international data on the subject, we will assume as a first approximation that the distribution of dependents by family income levels is not skewed. We can then discuss the volume of dependency expenditure in terms of the number of dependents in the population.

[15] The inclusion of the older age population as a factor reducing aggregate savings rates can be justified on two grounds. First, because of imperfect foresight (taken in conjunction with such conditions as inflation and imperfect capital markets, increasing life span, and rising "minimum" standard of living) they may not have made adequate provision for their nonproductive years, and hence they constitute a drain on the income and discretionary expenditure of others. Secondly, they may be running down their stock of savings so that their ratio of current savings to income is very low. Hence, at any point in time, *ceteris paribus*, a large proportion of retired people in the population would mean a lower aggregate savings ratio.

$$
\begin{array}{lll}
(3)\quad \ln S/Y = & 7.7762 + & .1037 \ln Y/N \\
 & (5.7914) & (2.3132) \\
 & — & .2437 \\
 & + .0234 \ln g - & 1.4893 \ln D_3 \\
 & (3.0428) & (4.8493) \\
 & .2387 & .5264
\end{array}
$$

$$\bar{R}^2 = .5729 \qquad F = 33.6357$$

The results support the hypothesis concerning the effects of dependency ratios on savings rates. The regression coefficients, which in these double logarithmic equations can be interpreted as elasticities, are much larger for D_1 and D_3 than for the other terms; and the *beta* coefficients indicate that the D terms contribute more of the statistical explanation. The t-values are all significant above the .025 level, indicating that the dependency variables should be considered statistically distinct variables in the explanation of aggregate savings rates.

$$
\begin{array}{lll}
(4)\quad \ln S/N = & 2.7851 + & 1.1486 \ln Y/N \\
 & (2.2571) & (21.5100) \\
 & + .0265 \ln g - & 1.3438 \ln D_1 \\
 & (3.5592) & (4.7916) \\
 & - .3966 \ln D_2 & \\
 & (2.6461) &
\end{array}
$$

$$\bar{R}^2 = .9559 \qquad F = 396.677$$

$$
\begin{array}{lll}
(5)\quad \ln S/N = & 3.2383 + & 1.0924 \ln Y/N \\
 & (2.5082) & (25.3417) \\
 & + .0245 \ln g - & 1.4857 \ln D_3 \\
 & (3.3327) & (5.0310)
\end{array}
$$

$$\bar{R}^2 = .9563 \qquad F = 533.913$$

Equations (4) and (5), which were estimated with the level of per capita savings as the dependent variable, gave similar results as well as some additional information. The relatively large elasticities of D_1 and D_3 and their high level of statistical significance suggest that the studies of savings levels which have shown the overwhelming importance of per capita income are in part picking up the effect of unspecified demographic variables. Another interesting result is the value of the regression coefficient on the per capita income term, which indicates an income elasticity of savings fairly close to unity. This is, of course, in close accordance with the general absence of an upward trend in S/Y ratios in underdeveloped countries over the medium term that we noted in the introduction. The regression coefficients of equations (2)–(5) indicate that, while large income differences (say, between \$100 and \$1,000 per capita) are associated with appreciable differences in savings rates (in the example given, 100 percent), the effects of higher income appear to be considerably reduced when dependency variables are included in the analysis.

IV. *Further Discussion and Results*

In interpreting these results, however, we should note some possible problems. First, multicollinearity. The simple correlations between $\ln Y/N$ and $\ln D_1$, $\ln D_2$, and $\ln D_3$, are respectively, $-.74$, $.80$, $-.67$. These are not as low as might be desired, but as the t-ratios indicate, the regression coefficients for the per capita income and the dependency variables were highly significant in all cases when they were included together. Furthermore, as noted below, equations estimated for different subsets of countries gave similar results. With different sample coverage and with lower intercorrelation between the independent variables, those equations, too, show large elasticities and a high level of statistical significance for the dependency variables.

There is also an identification problem. Can it be assumed that dependency rates are exogenous factors which influence savings rates, or, might not both variables be endogenous, determined perhaps by per

TABLE 1—PARAMETER ESTIMATES FOR SAVINGS IN 47 UNDERDEVELOPED COUNTRIES AND 20 WESTERN DEVELOPED COUNTRIES
(t-values are in parentheses)

Dependent Variable	Independent Variable				$\bar{R}^2$	F
	$\ln Y/N$	$\ln g$	$\ln D_1$	$\ln D_2$		
Underdeveloped countries						
1) $\ln S/Y$	.1292*	.0227**	−1.2297**	−.4455**	.2419	4.6685
	(1.8487)	(2.8079)	(2.7636)	(2.1554)		
2) $\ln S/N$	1.1167**	.0239**	−1.3122**	−.4469**	.8975	101.6727
	(16.8355)	(3.1204)	(2.9400)	(2.2783)		
Western developed countries						
3) $\ln S/Y$	.0035	.2589*	−.4324**	−.4916**	.4395	4.7245
	(.0296)	(1.6228)	(1.7099)	(2.6547)		
4) $\ln S/N$	1.0049**	.2591*	−.4300*	−.4914**	.8413	26.1798
	(8.3684)	(1.6208)	(1.6966)	(2.6477)		

* significant above the .1 level
** significant above the .01 level

capita income? For example, does the relation on which we have focused link savings behavior and dependency rates, or have we not simply related two of the many elements which constitute the complex of underdevelopment. In this perspective, any of these other conditions, e.g., low protein consumption, could perform equally well in providing a statistical "explanation" for savings rates.[16]

It is not clear what can be done to resolve this identification problem within the framework of a single equation model, to which the present work, like other studies of international savings behavior, has been confined. However, by partitioning the data and fitting separate equations for the underdeveloped countries and the Western developed countries, we can see whether or not the earlier equations did not simply reflect the high correlation between two of the interrelated features of underdevelopment. The parameter estimates for these subset equations are presented in Table 1.[17]

The results indicate differences between the two sets of countries. In fact, application of the Chow test of covariance [4] indicated that the developed and the underdeveloped country savings behavior belong to different regression structures. Within *both* groups of countries, however, the relation on which we have focused remains true. The dependency variables are statistically significant, and their regression coefficients indicate that they are quanti-

[16] The same objection can be made with equal validity to econometric models which use per capita income as an explanatory variable; for in this view, low income would be simply one of the many interrelated features of "underdevelopment." As the example of income indicates, however, there is an important difference if the variable used has also been selected on analytical grounds which make a plausible theoretical connection between the variables. As set out in Section II, this criterion is fulfilled in the case of dependency ratios and saving.

[17] The question arose as to whether to include the 7 Communist countries in the subset of the Western developed countries. Application of the Chow test of covariance [4] led to a (borderline) rejection of the null hypothesis; but because there were too few degrees of freedom to permit very meaningful reliance on the test and because of doubts concerning the comparability of their data, I did not include them.

tatively more important than the income terms as a determinant of aggregate savings ratios.[18] In the equations for the level of per capita savings, the regression coefficient on the per capita income term continues to indicate an income elasticity close to unity, slightly above for the underdeveloped countries and almost exactly one for the developed countries. Thus the subset results confirm the general pattern of savings behavior suggested by our hypothesis.

Considering how crude the data [14] and methodology are, these results can be regarded as providing support for the notion that birth rates and dependency factors are an important determinant of aggregate savings ratios. We should also note that the model we have used does not do full justice to the hypothesis it attempted to test. The independent variables relate to the supply of savings. Because of the restrictions imposed by the available data, however, the dependent variables refer to *ex post* savings and investment. This limitation is especially serious since the hypothesis concerning the burden of dependency charges on discretionary expenditure relates to savings *potential*. Conceivably, actual savings could have been higher in many countries with more favorable demand conditions.

V. *Analytical Implications*

The results were reached with the fairly simple reasoning outlined earlier, but they clearly have implications for the theory of savings. One way to conceptualize the findings would be to posit a model in which individuals maximize a utility which is a function of consumption (present and future), and of children. Considering the limited availability of contraceptive knowledge and devices in most underdeveloped and in some developed countries during this period, however, this might not be realistic; for technical rigidities (human physiology) may have interfered with optimizing behavior.[19] An alternative approach would be simply to integrate these findings with current theories of saving; for example, formulating a model of maximization over lifetime consumption, subject to the constraints imposed by expenditures for dependents.

In terms of their analytical implications, the cross section results are also consistent with the hypothesis that favorable demographic conditions have been one of the important factors which have made possible the high savings ratios of the Eastern European communist countries.[20] Since birth rates in those countries (with the exception of the USSR) had reached relatively low levels before the communists took power [31, pp. 276–79], [22], [8], this raises a question as to whether the special institutional features of these regimes are a sufficient condition for achieving high savings rates in underdeveloped countries.

[18] Concerning the multicollinearity problem mentioned earlier: for the 47 underdeveloped countries, the simple correlations between $\ln Y/N$ and $\ln D_1$ and $\ln D_2$ are, respectively, $-.42$ and $.52$; and for the 20 developed countries, $.11$ and $-.03$.

[19] For an economic analysis of fertility behavior assuming choice, see Gary Becker in [9]. Of course, it is not sure that high birth rates are due exclusively to lack of knowledge and access to contraceptive techniques. Paul Schultz' study of birth rates in Puerto Rico [24] provides some support for the hypothesis that larger family size in rural and premodern societies may be a rational response to the economic returns of parenthood in such environments.

[20] I say "one" because in equation (2), the residuals for the communist countries were not randomly distributed. The actual observations for 6 of the 7 communist countries were above the regression surface, with fractional residuals averaging 24 percent. A separate equation was estimated for the communist countries, and the Chow test was applied to see if these countries belonged, statistically speaking, to the same universe which generated the slope of the equation for the entire set. The test was not conclusive. At the 5 percent level, it was possible to reject the null hypothesis, but not at the 1 percent level of significance.

VI. *Policy Implications*

This paper suggests that high rates of population growth constitute a two-fold barrier to economic development. On the one hand, as often noted, a growing population increases the volume of development expenditure needed to improve the quality of the labor force and to raise capital-labor ratios. Our results indicate that through their effects on dependency ratios, high birth rates also limit an economy's ability to respond to these additional needs for savings and investment.

Our results relate only to the effect of the dependency burden on physical capital formation and measured saving, but large family size also affects human capital formation. Because conventional morality usually dictates that priority be given to maintaining the children's physical survival, sufficient resources may not be available for improving the quality of human resources by investments in training and education. Thus a large dependency burden may lead to an economically sub-optimal abundance of raw labor, relative both to skilled manpower and to the physical capital stock. Such a situation is, in fact, recounted of many underdeveloped countries.

Our discussion suggests that high birth rates and dependency ratios in the conditions of most underdeveloped countries may entail a sub-optimal allocation of resources due to physiological and institutional rigidities. That is, children are born to parents who might prefer not to have them born, but who are thereafter committed to supporting them. These dependents absorb a large portion of the resources potentially available for increasing the stock of physical and of human capital. The implications of these demographic conditions are even more distressing because current trends in many underdeveloped countries appear to be leading to higher rather than to lower dependency ratios. Birth rates in most underdeveloped countries have increased rather than fallen in the past decades [6, pp. 27–28] [31]. The dependency problem is further aggravated because when mortality rates fall in underdeveloped countries, the greatest beneficiaries are the youngest age cohorts. This reduced infant and child mortality adds to the dependency burden, and constrains still further the economy's saving potential.

If these conclusions are interpreted to warrant expanded policy efforts to lower birth rates in underdeveloped countries, two cases should be distinguished. In some areas, notably cities in transitional societies, the motivation for fewer births is present, as evidenced in high rates of crude induced abortion. Here, programs to make available less costly and painful birth-reducing techniques should be sufficient, though the scale and scope of the requisite effort may have to be much greater than is contemplated in current population programs [7]. In other areas, however, the problem may lie in creating the conditions for new attitudes toward choice in family size[21] and in reducing the technical rigidities which may inhibit rational choice in this area.

VII. *Conclusions*

This paper has provided statistical support for the suggestion, advanced by a number of economists, that demographic conditions are a major determinant of aggregate savings rates. Our results indicate that dependency ratios are a statistically distinct and quantitatively important influence on aggregate savings ratios, both

[21] For example, research by Irma Adelman indicates a relatively high elasticity of birth rates with respect to education, especially in underdeveloped countries [1, pp. 321–23, 336–37]. These results support the suggestions which have been made on other grounds, on a high policy priority for rural literacy programs.

for the 74 countries considered as a whole and within the subsets of developed and underdeveloped countries. Introduction of dependency variables also pointed to a specification error present in other cross-section studies of savings levels. Considerable work obviously remains to be done, however, on international savings behavior.[22]

High dependency ratios—and ultimately high birth rates—are among the important factors which account for the great disparity in aggregate savings rates between developed and underdeveloped countries. Cross-section analysis showed an income elasticity of savings not much above unity. These results, and the stability (if not upward trend) in birth and dependency rates, help explain the failure of aggregate savings ratios to rise with increasing income levels in most underdeveloped countries.[23]

Our analysis indicates pessimism concerning the possibility of achieving substantial increases in the savings rates of underdeveloped countries unless birth rates are reduced. The findings also indicate a medium term rigidity in savings rates for it will be several years before the dependency ratio D_1 will fall sharply even if birth rates are brought down. On the latter score, as noted, there are no grounds for optimism. Particularly in some less developed countries, government birth control programs have often been hampered by the opposition of the intellectuals and of elite opinion, including many economists, on nationalist grounds. The present results may be relevant in demonstrating that by limiting their capacity for domestic saving and fixed capital formation—goals that are widely desired—high birth rates may conflict with their own development aims.

[22] Even within the framework of single equation models, presently available data do not permit specification of such relevant variables as rates of return, efficiency of capital markets, size distribution of income, and time preference for most countries.

[23] This question is treated in a very different but complementary manner in [18].

Appendix:

Data Sources and Methodology

National income and demographic data were available for the following 74 countries: Costa Rica, El Salvador, Puerto Rico, Honduras, Nicaragua, Dominican Republic, Jamaica, Trinidad and Tobago, Panama, Argentina, Brazil, Chile, Ecuador, Mexico, Paraguay, Peru, Venezuela, Uruguay, Malta, Cyprus, Greece, Spain, Portugal, Turkey, Iran, Barbados, British Guiana, Egypt, Tunisia, Uganda, Mauritius, Tanganyika, Ghana, Sudan, Kenya, Iraq, Jordan, Israel, Morocco, Ceylon, India, Pakistan, Thailand, Taiwan, South Korea, Malaysia, Philippines, Japan, Belgium, Luxembourg, France, West Germany, Italy, Netherlands, Denmark, Iceland, Norway, Sweden, Finland, Austria, Eire, United Kingdom, Switzerland, Australia, New Zealand, Canada, United States, USSR, Hungary, Bulgaria, Czechoslovakia, Yugoslavia, East Germany, and Poland.

Data on population and dependency ratios were computed from the United Nations [31]. 1964 data on Gross National Product (GNP) and aggregate domestic savings in 40 underdeveloped countries were made available by the Statistics and Reports Division of the Agency for International Development. Gross domestic savings was measured as the difference between gross investment and the current account balance. Local currency figures were converted at official exchange rates, except in the cases where the country desks at *A.I.D.* judged that another rate, based on purchasing power parity criteria, was more appropriate. For the remaining noncommunist countries, data on savings and Gross Domestic Product (*GDP*) were taken from United Nations [35]. For the developed countries, the data of Table 9A, which converted local currencies at official exchange rates, were used. For the underdeveloped countries, Table 9B, which employed purchasing-power parity rates for the conversion, was used. Data on the trend rates of

national income growth were taken from [34] and [37]. For the communist countries estimates of *GDP* at factor cost and its growth were taken from Stanley H. Cohen and Maurice Ernst [36]; and from Frederick Pryor and George Staller [23]. Figures for the share of Gross Investment in *GNP* at factor cost for Poland, Hungary, Czechoslovakia and the USSR are from Thad Alton [2, p. 92]; Bulgaria [36]; East Germany, from Wolfgang Stolper [27, p. 437]; and for Yugoslavia, at market prices, from Table A-35 in [32].

REFERENCES

1. I. ADELMAN, "An Econometric Analysis of Population Growth," *Amer. Econ. Rev.*, June 1963, *53*, pp. 314–39.
2. T. P. ALTON, et al, *Polish National Income and Product in 1954, 1955 and 1966.* New York 1965.
3. H. B. CHENERY AND P. ECKSTEIN, "Development Alternatives for Latin America," *Project for Quantitative Research on Economic Development*, Harvard Univ. 1967.
4. G. C. CHOW, "Tests of Equality Between Sets of Coefficients in Two Linear Regressions," *Econometrica*, July 1960, *28*, 591–605.
5. A. J. COALE, "How the Age Distribution of a Human Population is Determined," *Cold Spring Harbor Symposia on Quantitative Biology*. 1957.
6. O. A. COLLVER, *Birth Rates in Latin America*. Berkeley 1965.
7. K. DAVIS, "Population Policy: Will Current Programs Succeed?", *Science*, No. 3802, Nov. 10, 1967, 730–40.
8. P. DEMENY, "Early Fertility Decline in Austria-Hungary: A Lesson in Demographic Transition," *Daedalus*, spring 1968, 502–22.
9. *Demographic Change in Developed Countries*, Nat. Bur. Econ. Research. Princeton 1960.
10. J. S. DUESENBERRY, *Income, Saving, and The Theory of Consumer Behavior*. Cambridge 1949.
11. W. EIZENGA, *Demographic Factors and Savings*. Amsterdam 1961.
12. M. FRIEDMAN, *A Theory of The Consumption Function*. Princeton 1957.
13. A. HIRSCHMAN, *The Strategy of Economic Development*, New Haven 1958.
14. R. HOOLEY, "The Measurement of Capital Formation in Underdeveloped Countries," *Rev. Econ. Statist.*, May 1967, *49*, 199–208.
15. H. S. HOUTHAKKER, "On Some Determinants of Savings in the Developed and Underdeveloped Countries," in *Problems of Economic Development*, E. A. G. Robinson ed., London 1965.
16. S. KUZNETS, "Quantitative Aspects of the Economic Growth of Nations: Capital Formation Proportions," *Econ. Develop. Cult. Change*, 8, No. 4, Part II, July 1960.
17. L. LANDAU, "Determinants of Savings in Latin America," *Project for Quantitative Research on Economic Development*, Harvard Univ., 1966.
18. N. H. LEFF, "Marginal Savings Rates in the Development Process: The Brazilian Experience," *Econ. J.*, Sept. 1968, *78*, 610–23.
19. T. MIZOGUCHI, "International Comparison of Savings Ratios by Occupational Groups," *Hitotsubashi J. Econ.*, Feb. 1967, 61–74.
20. F. MODIGLIANI, "The Life Cycle Hypothesis of Saving, the Demand for Wealth and the Supply of Capital," paper presented to the Rome Congress of the Econometric Society, 1965.
21. ———, AND R. E. BRUMBERG, "Utility Analysis and Aggregate Consumption Functions," mimeo 1953.
22. *Population Trends in Eastern Europe, the USSR and Mainland China*, Milbank Memorial Fund. New York 1960.
23. F. L. PRYOR AND G. STALLER, "Dollar Value of the Gross National Product in Eastern Europe, 1955" *Econ. Planning*, No. 1, Feb. 1966, 1–27.
24. T. P. SCHULTZ, "A Family Planning Hypothesis: Some Empirical Evidence from Puerto Rico," The RAND Corporation 1967.
25. H. W. SINGER, "Population and Economic Development," paper presented

to the 1954 World Population Conference; reprinted in *International Development: Growth and Change.* New York 1964.
26. J. J. Spengler, "The Population Obstacle to Economic Betterment," *Amer. Econ. Rev.*, May 1951, *41*, 343–54.
27. W. Stolper, *The Structure of The East German Economy.* Cambridge 1960.
28. A. M. Strout, "Savings, Imports, and Capital Productivity in Developing Countries," paper presented to the Rome Congress of the Econometric Society, 1965.
29. S. Swamy, "A Dynamic, Personal, Savings Function and Its Long-Run Implications," *Rev. Econ. Statist.*, Feb. 1968, *50*, 111–16.
30. Charles Wolf, Jr., "Savings Regressions, 'Self-Help' and Development Performance," The RAND Corporation, Santa Monica 1964, pp. 1–28.
31. United Nations, *Demographic Yearbook 1965.* New York 1966.
32. ———, Economic Commission for Europe, *Some Factors in Economic Growth in Europe During the 1950's.* Geneva 1964.
33. ———, *Economic Survey of Latin America*, 1963. New York 1965.
34. ———, "The Financing of Economic Development," *World Economic Survey, 1965.* Part I. New York 1966.
35. ———, *Yearbook of National Income Accounts Statistics, 1965.* New York 1966
36. U.S. Congress, Joint Econ. Com., *New Directions in The Soviet Economy.* Washington 1966.
37. U.S. Agency for International Development, "Gross National Product: Growth Rates and Trend Data," March 1967.

COMMUNICATIONS

Fiscal and Monetary Policy Reconsidered

By ROBERT EISNER*

Does the 1968–69 period of continued inflation in the U.S. economy in the face of higher tax rates and a huge swing from budget deficit to surplus prove that fiscal policy is impotent? One-third of a century after Keynes' *General Theory* are we to return to the policy prescriptions of a crude—or sophisticated—quantity theory of money? Is salvation yet to be found in a restrictive monetary policy which has sharply reduced the rate of growth of the money supply to some 3 percent per annum, seen bank prime rates rise to 8-1/2 percent, rates on top corporate bonds to over 8 percent, and average mortgage yields to 7-3/4 percent?[1] All this coincides with relatively full employment—a seasonally adjusted unemployment rate of 3.6 percent in June 1969—and, after over a year of the 10 percent income tax surcharge and perhaps nine months of tight money, accelerating price inflation now approaching an annual rate of 6 or 7 percent.

We shall argue in this paper that contemporary economic theory implies severe limitations in the effectiveness of both tax policy and monetary policy in curbing an inflation brought about by high government expenditures, such as are typically related to war. In so doing, we shall review the balanced budget theorem in light of the permanent income hypothesis. We shall also review monetary theory in terms of interaction of wealth and liquidity factors in current demand.

I. *Weakness of Compensatory Tax Policy*

The balanced budget theorem indicated that, with appropriate *celeris paribus* assumptions, an increase in government expenditures for goods and services matched by a corresponding increase in tax revenues will under certain circumstances imply an increase in the new equilibrium level of national income equal to the increase in government expenditures (and taxes). A ready implication of the theorem is that an increase in government expenditures at full employment would have to be matched by a somewhat greater than equivalent increase in tax revenues, or a tightening of monetary policy, if inflation is to be avoided. Conversely, of course, to maintain full employment in the face of a fall in government demand for goods and services, there must be a greater than equivalent reduction in tax revenues and/or an easing of monetary policy. Applying a very simple example of the balanced budget relation to the problem of maintaining aggregate demand constant, since the multiplier effect of an increase in government expenditures for goods and services is $1/1-b$ while that of an increase in taxes is $-b/1-b$, the increase in taxes necessary to counterbalance a one dollar increase in government expenditures must be $1/b$, where b is, of course, the marginal propensity to consume.

Inflationary ills of the last several years in the U.S. economy are, consistent with the analysis above, frequently traced to the major escalation of Vietnam war expendi-

* Professor of economics, Northwestern University, and research associate, National Bureau of Economic Research. The author is most grateful for comments by the Editor, Frank Brechling, Otto Eckstein, Walter D. Fisher, Milton Friedman, Sherman J. Maisel, and Charles L. Schultze on an earlier draft of this paper. They are, of course, all absolved of responsibility for any of the current version with which they are in disagreement.

[1] The Federal Reserve Bank of New York reported in August 1969, "... While there are suggestions of slowing in some sectors of the economy, there is little evidence that financial restraint has as yet sufficiently dampened the excessive pace of economic activity." [11, p. 155]

tures beginning in 1965, unmatched by a reasonably prompt tax rise. Over the years from 1965 to 1968, indeed, Federal expenditures for national defense rose from $50.1 billion to $60.6 billion to $72.4 billion to $78.9 billion, or from 8.6 percent of 1965 Gross National Product (*GNP*) to 9.2 percent of an already inflated 1968 *GNP*. The surcharge of 1968 is estimated to have amounted to somewhat more than $6 billion in personal income taxes[2] and probably came to another $4 billion in increased corporate income taxes plus a few billion more in the accompanying extension of certain excise rates. Operating on the apparent assumption that much of the increase in defense expenditures had already been absorbed in the real and inflationary growth of the economy, some observers ascribed the failure of the tax increases of 1968 to curb further inflation to a premature move by the Federal Reserve System to easy money, which counteracted what would otherwise have proved an effective, if belated, anti-inflationary fiscal policy. It is the thesis of this paper that the tax surcharge should never, on basic theoretical grounds, have been considered an effective anti-inflationary device and that, given a sufficiently excessive rate of government spending, there is little that any meaningful monetary policy can do to stop inflation.

The basic economic error of those who saw in the 10 percent income tax surcharge an adequate measure against inflation may be charged to failure to take into account the implications of the permanent income hypothesis, both for consumption and, in correlative fashion, for business investment. Dealing in terms of equilibrium or permanent income, an increase in tax withdrawals of some ten or eleven billion dollars per year may be expected to have a very substantial depressant effect on aggregate demand, in the order of perhaps $8 billion on consumption alone. But the tax increase legislated in the surcharge did not represent a corresponding reduction in *permanent* after-tax income. It did clearly represent a reduction, in itself, in *current* or measured after-tax income for roughly a twelve month period. Before we can make any meaningful predictions of the effect of this on consumption, we must note its relation to permanent income, or more precisely, to a fully specified consumption function.

To do so, let us invoke a simplified version of the Modigliani-Brumberg formulation [7] and write consumption of any period as

$$C = C(Y, Y^e, W), \tag{1}$$

where Y is current income, Y^e some appropriately weighted and discounted average of expected future incomes, and W is initial wealth. For the average individual, with a number of years of income-earning and probably a somewhat larger number of years of life in prospect and, frequently, some substantial body of wealth already accumulated, current income is the least important argument of this function. The notion that current income has much at all to do with current consumption is probably rooted in two confusions, one theoretical and the other empirical. The theoretical confusion involves the tendency to carry over reduced forms of a system of equilibrium relations to cyclical and dynamic problems where they do not apply. Thus, if we were to compare two states, for an individual household or for an economy, in which expected income and assets were in a fixed and equilibrium relation to current income, we should properly expect consumption in these two states to differ by a very high proportion of the differences in current income. But there is no good reason to think that the two states of comparison, as an economy before and after surtax, should involve such an equilibrium relation among the relevant variables. And the complementary error in empirical analysis is to infer that because in some sets of data there is a fairly high "marginal propensity to consume" out of current income, there is any stable relation between consumption and current income that is not crucially dependent upon the relation involving current income and the other, omitted variables in the consumption function. This may be seen readily by examining the marginal propen-

[2] [10, p. 463].

sity to consume out of current income, derived from (1), which may be written:

$$\frac{dC}{dY} = \frac{\partial C}{\partial Y} + \left(\frac{\partial C}{\partial Y^{e}} \cdot \frac{dY^{e}}{dY}\right) + \left(\frac{\partial C}{\partial W} \cdot \frac{dW}{dY}\right). \tag{1.1}$$

The main point to be gleaned from (1.1) and the most casual empiricism is that the *partial* derivative of consumption with respect to current income, $\partial C/\partial Y$, even for income measured over a period as long as a year, is likely to be very small. Indeed, in the simplest Modigliani-Brumberg case, for a not atypical household with, say, 25 years of anticipated future existence, aside from discounting factors, it would be .04. Bringing in a positive rate of time discount would certainly increase this *mpc* above .04, but it may still be seen as likely to be very low. If a household's assets and expected income are unaffected, we can expect little in the way of change in consumption from a temporary reduction in income brought on by an increase in tax rates.

Yet, it is precisely this issue that has been largely ignored in discussions of likely effectiveness of the 1968–69 surcharge as a counterbalance to escalated Vietnam war expenditures. For the latter were very substantially a direct contribution to the domestic demand for goods and services. The tax increase would offset this demand via its indirect effect on consumption and investment demand. But its initial effect on consumption, we may note in (1.1), is likely to be very small. First, the contribution of the last term may reasonably be expected to be zero, on the ground that assets held by households should be no more after the imposition of the tax than before. In fact, if anything, the role of the last term turned out to be perverse, as the concurrent declines in interest rates entailed higher prices for outstanding bonds and equities. The influence of the second term may reasonably have been expected to be small, insofar as it applies to the direct or initial effect of taxes, if only because the tax increase was widely advertised as temporary, scheduled to end in little more than one year. Indeed, even apart from the legal status of the tax as temporary, individuals would hardly be expected quickly to adjust their long-run future income expectations in response to a tax change of this sort. The change in consumption due to a change in expected future income should not therefore have been foreseen as large. We are thus left with the small direct effect of the change in current after-tax income as the major component of what initial anti-inflationary impact on consumption might reasonably have been expected from the tax surcharge. Is it really any wonder that consumption continued to rise after the imposition of the surtax in 1968, while personal saving, the difference between current or measured after-tax income and consumption, dropped precipitously, from almost 7.5 percent of disposable income in 1967 and the first half of 1968, to 6.6 percent in the final half of 1968, and 6.1 percent in the first quarter of 1969?[3]

As all students of multiplier theory know, however, each initial "injection" of purchasing power generates further demand as secondary income recipients excercise their propensities to consume. While perceptive income recipients might reckon the induced changes in income which they experience as no more permanent than the original tax change itself, it is probably more likely that they will fail to distinguish tax-induced changes in income from changes unrelated to taxes. The propensity to consume which

[3] [10, p. 467]. It should be recognized that the Modigliani and Friedman models relate directly to the consumption of services and not to consumer expenditures, which include investment in consumer durables. A small change in the demand for consumer services might conceivably generate an accelerator effect entailing more substantial investment (or disinvestment) in the consumer goods which would provide these services. There may well be some magnification of the role of income tax changes along these lines. It may appear doubtful, however, that a small short-run change in the demand for consumer services would generate any great investment in the durable goods which would provide services for a long time. The arguments in favor of a permanent income theory for business investment [2] may reasonably apply also to consumer investment.

we may wish to apply, while still not an *mpc* out of permanent income, or one reflecting equilibrium relations of current income to wealth and expected future incomes, may hence perhaps be higher than that out of the initial tax-created change in after-tax income.

Modifying one of the formulations of William Salant [9] in his treatment of the balanced budget theorem, we note that the relative change in income resulting from the increase in a tax proportionate to income may be written

$$(2) \qquad \frac{Y_2 - Y_1}{Y_1} = - \frac{b_i(t_2 - t_1)}{1 - b_s(1 - t_2)},$$

where b_i is the marginal propensity to consume out of the initial, tax-generated change in disposable income, b_s is the marginal propensity to consume out of the secondary or subsequent rounds of changes in disposable income, t_1 and t_2 are the tax rates and Y_1 and Y_2 the income equilibria, before and after the tax rate changes, respectively. While somewhat inappropriately applying this equilibrium relation to the issue at hand, we may still recognize b_i as very small. If then, for example,[4] $b_i = .1$, $b_s = .5$, $t_1 = .18$ and $t_2 = .198$, the relative depressant effect of the surcharge on income would be

$$(2.1) \qquad \frac{Y_2 - Y_1}{Y_1} = \frac{.1(.198 - .18)}{1 - .5(1 - .198)} \cong .003$$

or three-tenths of one percent—less than $3 billion, even at our current inflated rate of *GNP*.

If the predictable restraining effect of the surcharge on consumption should have been small, it should have been smaller still on business investment. Some conventional wisdom might have argued that the surcharge on corporate income should have made business investment less profitable and hence reduce it. I have on a number of previous occasions argued against this view.[5] The major aggregative effect of profits on investment must relate to the expected profitability or marginal efficiency of investment. This surely involves expectations of a long-run future or "permanent" income or output and demand.[6] The momentary ripple in corporate profits after taxes should and did have little effect on business investment demand. And since the impact of the surcharge on consumption was negligible, as we have noted, little or no acceleration effect of the surcharge on investment was to be found either. To some extent, as business endeavors to smooth production, inventory investment might even have been expected to compensate for slight short-run fluctuations in consumption.[7]

Was there hence nothing that fiscal policy could have done to curb inflation? This does not follow, but the political constraints accepted by otherwise sophisticated advocates of fiscal policy in the economics profession led them to ignore or even oppose the possible fiscal measures which would have brought a speedy end to inflation. By doing so, they curiously left the field on the one hand to conservative opponents of government social programs and, on the other, to the current "I-told-you-so" band of monetary enthusiasts. For there were those insisting, as the surcharge was being debated, that they would support an increased tax rate only if it were accompanied by reduced government expenditures. To some others mesmerized by simplistic application of the lessons of Economics/1—before the permanent income theory took hold—this insistence seemed to contain the elements of heresy and blackmail. If the surcharge were set at an appropriate level, why have a reduction of government expenditures *in addition*? Would this not be "overkill?" Indeed, fears of "overkill" apparently swept enough of the fiscal-minded economic community in the waning months of the Johnson Administration for the Federal Reserve Board to be

[4] Basing figures other than b_i very roughly on equations (A.1) to (A.5) and on (A.34) in Fromm and Taubman [6, pp. 126–27 and p. 133].

[5] In [1], for example.

[6] See [2].

[7] The corporate tax surcharge may actually also stimulate business capital expenditures! Current investment entails current start-up costs and additional depreciation charges which are balanced against the expectation of added future income. Higher current tax rates which are not expected to persist would encourage current investment and charges against current highly taxed income in the expectation of future income subject to lower rates of taxation.

induced to ease monetary restraints shortly after the enactment of the surcharge in 1968.

Of course, as it turned out, the ceiling on Federal expenditures proved decidedly porous. It was inapplicable, to begin with, to the original source of inflationary pressure, expenditures for Vietnam. What with other "loopholes," Federal expenditures for goods and services continued to creep up all through 1968 and into 1969. (And state and local government expenditures increased at a faster rate.) Inflation therefore continued, partly as the lagged response to, and projection of, pressures and movements begun earlier, partly because government expenditures continued to increase, and in considerable part because the tax increase did little, and should have been expected to do little, to reduce private demand.

There was of course one fiscal measure which might have been expected to do the trick: a real and substantial reduction in the war and other defense spending which had contributed to the inflation in the first place.[8] There were a considerable number of economists who advocated this, and who went a step further, arguing that rejection of the surcharge would help force the hand of the Administration in the direction of reversing the prevailing escalation of the war. This was, of course, a political judgment, but not an implausible one to many of us then, and certainly not now in retrospect, in view of the pressures which were at the time leading President Johnson to proclaim a change in policy and his own withdrawal from the forthcoming presidential race. It may also be well to note that war and defense demand are likely to be more inflationary, dollar for dollar, than other components of aggregate demand. The nature of procurement practices and the psychological aura of necessity that accompanies military purchasing have undoubtedly led to acquisition of military goods at particularly inflated prices. Major military expenditures then have the effect of raising prices throughout the economy in the competitive bid for factors and materials fueled by firms with little pressure to keep down costs.

But the path of ending the war and/or drastically cutting military expenditures was not the one followed. Prices continued to rise, apparently at an accelerating rate! And with the failure of an improperly conceived fiscal policy has come a growing disenchantment with fiscal policy in general and new strength for the notion that it is money which really matters. We are now told that, if only the quantity of money can be kept from growing immoderately, inflationary forces will, with due lags, subside. This would occur, it is implied, regardless of fiscal policy, that is, regardless of the rate of Federal expenditures and, for that matter, taxation. Having noted the ineffectiveness of the tax surcharge in checking inflation, we have now to demonstrate that tight money, or a tight rein on the money supply in the face of substantial continued inflationary pressure from government expenditures, will be similarly ineffective.

II. *Inadequacy of Monetary Policy*

Let us elucidate the theoretical relations by constructing a simple model in which aggregate demand is the sum of government and private demand, and private demand is in turn a function of current after-tax income, expected future after-tax income, the rate of interest and privately held assets, including the real stock of money. The nominal stock of money may be taken, for simplicity, to be entirely Treasury currency; that is, an obligation of the government to the private sector of the economy, and a parameter to be fixed by the monetary authority. The demand for money will be taken as a function of the rate of interest, output, and assets. Income, assets and demand and its components are real. These

[8] It is curious that so many arguing for reductions in Federal spending have until very recently treated military expenditures as sacrosanct and concentrated their fire on nonmilitary or domestic programs. Yet in 1968, for example, those wishing to cut nonmilitary Federal expenditures for goods and services—and it is spending for goods and services which has a direct, full effect on aggregate demand—had only $21.1 billion at which to whittle away, while National Defense expenditures were $78.9 billion, almost four times as much. We may hence stress, on the basis of plain arithmetic, questions of priorities aside, that meaningful cuts in the direct Federal contribution to inflationary demand would have to involve cuts in war and defense spending.

relations and the conventional partial derivatives may be conveniently written as follows:

(3) Private Commodity Demand
$C = C(Y, Y^e, i, A, M)$
where $0 < C_Y, C_{Y^e}, C_A, C_M < 1, \; -\infty > C_i > 0$

(4) Real Money Demand
$M^D = M^D(X, i, A)$
where $M_X > 0, M_i < 0, M_A > 0$

(5) Output Equilibrium
$X = C + G$

(6) Money Market Equilibrium
$M = M^D$

(7) Money—Price Identity
$M = Q/P$

C = the sum of consumption and investment demand
Y = current income after taxes
Y^e = an appropriately weighted average of expected future income after taxes
i = the rate of interest or rate of time discount
A = the net real value of nonmonetary assets held by the private sector
M = real cash balances held by the private sector
M^D = real cash balances demanded by the private sector
X = output = income = Y + taxes
G = government demand for goods and services
Q = the nominal quantity of money
P = the commodity price index[9]

The model presented in equations (3)–(7) would appear to embody the essential considerations and parametric assumptions of fiscal and monetary enthusiasts alike. Yet, given this model, the limitations of monetary policy and the speciousness of the developing notion that *only* money matters may be made readily apparent.

In our model, the nominal quantity of money, Q, and the real rate of government demand for goods and services, G, are given. Let us then first ascertain whether an increase in government demand, measured by a corresponding increase in government expenditures for goods and services can be offset by maintenance of a constant stock of money. To further simplify the exposition, we will assume initially that the increase in government expenditures is unmatched by an increase in tax revenues, although more generally we might merely assume initially that the increase in government expenditures is less than fully matched by an increase in tax revenues. Such deficit-financed government expenditures must imply an increase in the quantity of money, Q, an increase in interest-bearing Treasury obligations, or both. Let us assume that the deficit financing is covered by direct borrowing from the public so that the quantity of money remains constant while private holdings of interest-bearing Treasury obligations increase. If we further assume full employment, there can be no increase in output, X, which means that the new equilibrium must involve a lower level of private commodity demand, C. But with tax revenues unchanged, an unchanged full employment level of output implies that current real income after taxes, Y, must also be unchanged. There is no a priori reason to assume any change in expected income, Y^e. Hence, this means that the reduction in private commodity demand must be associated with changes in the rate of interest, i, the real value of nonmonetary assets, A, or real cash balances, M.

Now, as long as the government budget deficit persists, unless prices or interest rates rise, A must rise as a consequence of the increase in interest-bearing Treasury obligations held by the private sector.[10] As indicated in (3), this in itself would raise commodity demand. It follows that the reduced commodity demand required by the new equilibrium would have to be associated with an increase in the rate of interest, i, which would also serve to lower A, and/or a reduction in M brought about by an increase in the level of prices. This last would further

[9] It will be noted below that in this system of five equations which is not explicitly closed, C, i, M, M^D, and P are endogenous. Y, Y^e, X, G, and Q are exogenous, either fixed or determined by relations not set forth above. A is partly exogenous and partly dependent on other variables in the system.

[10] The argument advanced in some quarters that Treasury borrowing may only prove a substitute for business or other private borrowing is irrelevant in that we are concerned here with the *net* value of nonmonetary assets held by the private sector. We join Patinkin [8], for example, and others in ruling out distribution effects of assets and liabilities so that if Paul is given an I.O.U. by Peter, thus increasing Paul's net assets but decreasing Peter's by an equal amount, we assume no resultant change in private commodity demand.

serve to reduce or even reverse the increase in A.

If the reduction in private commodity demand is associated with an increase in prices which reduces M and, possibly A as well, our point has already been demonstrated. The increase in government expenditures, with the nominal quantity of money kept constant, has forced a new equilibrium level of prices higher than the old. But the result would be the same even if no increase in prices is generated directly in the commodity demand equation. For the increase in the rate of interest, which may be looked at from the securities side as resulting from the increase in supply of Treasury interest-bearing obligations, or from the real side as implicit in the bidding of resources away from private investment to production for government, must have its consequences, in any event, in the money market. If the increase in the rate of interest is sufficient to prevent the value of nonmonetary assets from rising, it is immediately clear from (4) that the real demand for money must be less. This means that, to satisfy (6), in accordance with the constraint shown in (7), the real supply of money, M, must be reduced by an increase in prices, P. If the value of nonmonetary assets has risen because the net increase in Treasury obligations corresponding to the budget deficit outweighs losses in capital values due to higher interest rates, it would still seem clear that the real demand for money must fall. To claim otherwise would be to argue that the higher rate of interest necessary to induce investors to move from money to bonds had somehow left them content to hold more money than initially. Recognizing the interdependence of markets, it would be to argue that the increase in interest rates necessary to reduce commodity demand despite the increase in the net value of assets would not be sufficient to reduce the real demand for money, implying a shift from commodities to money as interest rates and bond holdings rise. Indeed, as will be seen below, this would be to to argue that a deficit-financed increase in government expenditures can be deflationary while the same increase in government expenditures would be inflationary if fully supported by increased taxes.

The general argument which we have just developed follows through on the assumption that the increase in government demand is fully financed by taxes. If the increased taxes are taken to be temporary, it is clear that since, according to (3), $C_Y<1$, an increase in tax withdrawals equivalent to the increase in government expenditures does not reduce private commodity demand by as much as government demand has increased. To preserve output equilibrium (5), therefore, with no increase in Treasury interest-bearing obligations which would raise A, there must be an increase in the rate of interest, i, or a decrease in real cash balances, M. With the nominal quantity of money, Q, kept constant, real cash balances can be reduced only by an increase in the price level. And if we were to argue that the necessary reduction in private commodity demand could be accomplished entirely by an increase in the rate of interest, along with a concomitant reduction in the value of assets, we would still be faced with repercussions in the money market. Real money demand would now be reduced by both interest and asset effects, as indicated in (4), and to preserve the equilibrium condition (6), real cash balances would have to be reduced, in accordance with (7), by an increase in the price level.[11]

III. *Conclusion*

It should thus be clear that an increase in government demand for goods and services will prove inflationary with an equivalent increase in current tax revenues and the prevention of growth in the money supply.[12]

[11] It may be added that even if the tax increase is expected to be permanent we may have inflationary pressure in the instance above. For unless the sum of the first two derivatives in (1.1), that is, the marginal propensity to consume out of current and expected future income, is equal to or greater than unity, hardly generally likely in view of the role of assets, an increase in government expenditures matched by a permanent increase in tax revenues will require a downward shift in investment demand, or a higher price level, to restore equilibrium.

[12] How inflationary—that is, magnitudes as opposed to mere directions of effects— is not indicated by the

It is of course conceivable that a greater than equivalent increase in tax revenues and an actual reduction of the money supply might be sufficient to offset the increase in government demand. But as I have argued elsewhere [3], monetary measures are likely ultimately to be as limited in their impact in combating inflation as they have long been recognized to be limited in combating a deep depression. There may well be a "liquidity leak" of money substitutes, in the face of an attempt to stifle inflation by traditional monetary controls, akin to the liquidity trap facing those who would use monetary policy to combat a depression.[13] Yet we must note that conventional countercyclical fiscal policy will also prove severely limited precisely to the extent that it is recognized as counter-cyclical. For the permanent income theory should have forewarned us that the usual attempts to manipulate demand by varying taxes on income, as with the tax surcharge, would require fluctuations in tax rates far beyond those usually considered. These, in turn, given our difficulty in knowing precisely the magnitude as well as the timing of consequences of tax rate alterations, might well lead to less rather than more stability.[14]

All this should not leave matters entirely hopeless, at least in terms of fiscal policy. For one thing, as we have stressed, variation in government expenditures for goods and services remains a direct and highly potent factor, for good or for evil. Public works or, more generally, government investment and consumption, reemerge in their early role as prime weapons in the arsenal against depression and take on analogous importance in any struggle against inflation. If the government begins to contribute to inflation by rapidly escalating expenditures for a war such as in Vietnam, the indicated fiscal measures would include: 1) cut government expenditures for nonmilitary goods and services (to the extent this is feasible), as conservatives are generally quick to advocate; 2) cut other military expenditures—if we are to spend more in fighting a current war, reduce investment in strategic weapons for some future war; 3) recognizing the true, unavoidable cost to the economy of a Vietnam-type escalation, nip it in the bud, and the inflation along with it.

Further, however, there are other fiscal measures which, at the cost of varying degrees of distortion in the economy, might alter private demand sufficiently. On consumption, for example, there might be a set of very substantial sales taxes and/or subsidies which would be known to vary over time. Thus, there might be a very high excise tax attached to certain consumer goods and services, particularly those more easily postponable or for other reasons with a high elasticity of demand, with clear and widely publicized information that such rates would be changed as economic conditions change. One might introduce a 50 percent additional tax for buying a new automobile now with the suggestion that such tax would be removed a year from now and even that it

formal analysis of this paper. Conceivably, if the absolute values of C_i and C_M or, better, the elasticities of commodity demand with respect to the rate of interest and real cash balances, are very large, there need be little change in the rate of interest or in the price level in order to reduce commodity demand in (3) by the amount of increased government expenditures. Then further, if the absolute value of M_i, or the interest elasticity of the real demand for money, is very low, the small change in the rate of interest emanating from (3) would require little in the way of increases in the price level to reduce the supply of money and restore equilibrium in the money market. If commodity demand were perfectly interest-elastic—a kind of "investment trap"—no change in prices would be involved but commodity demand itself would be indeterminate over the range of perfect elasticity. Similarly if we have zero elasticity of the demand for money, depending on asset effects, there might be no price movement necessary to restore equilibrium. In the obviously realistic intermediate cases, relative potency of fiscal and monetary policies may then be seen to hinge on relative values of the interest and asset elasticities of the demands for commodities and the demand for money. Monetary enthusiasts must, implicitly or explicitly, see the commodity demand elasticities as high and the money demand elasticities, at least that with respect to the interest rate, as low.

[13] In addition, we should be aware that higher costs of money, like sales and excise taxes and, in the long run, business income taxes as well, tend to raise prices by raising costs of production.

[14] See the fine discussion of this matter by Milton Friedman [5].

might be replaced with a 20 percent subsidy. Indeed, promised future subsidies might be used generously as a desirable alternative to current excise taxes which would combat inflation by actually raising current prices to the ultimate consumer. Better still, there might be substantial delayed tax or interest benefits made available on nonnegotiable Treasury saving bonds.

Similar measures directed at inter-temporal substitution may also be applied to investment. Thus, a system of investment tax credits and investment allowances or subsidies with particularly high marginal impact, for example, in the order of 25 percent on investment above 80 percent of a past average rather than 7 percent on all investment, and also known to be variable over time, should have some very substantial effect on this component of private demand. A business firm might well delay purchase of equipment or construction during an inflationary situation in which it would have to pay an added tax of 25 percent on the investment while anticipating that the same investment a year or two from now would have no tax or perhaps a subsidy of 25 percent.

It would seem important for those attempting to manage the economy to learn some of these lessons, at least implicit in recent economic theory, before, for want of correct analysis and prescriptions, the "dismal science" falls into dismal repute, and our fates are turned back to know-nothings, within or without the profession.

REFERENCES

1. R. EISNER, "Capital Expenditures, Profits and the Acceleration Principle," in *Models of Income Determination*, Nat. Bur. Econ. Res., Stud. In Income and Wealth, Vol. 28, Princeton 1964.
2. ———, "A Permanent Income Theory for Investment: Some Empirical Explorations," *Amer. Econ. Rev.*, June 1967, *57*, 363–90.
3. ———, "Factors Affecting the Level of Interest Rates: Part II" (orginally "The Liquidity Trap and the Liquidity Leak") in *Conference on Savings and Residential Financing, 1968 Proceedings of United States Savings and Loan League*, Chicago 1968, pp. 28–40.
4. M. FRIEDMAN, *A Theory of the Consumption Function*. Princeton 1957.
5. ———, "The Effects of a Full Employment Policy on Economic Stability: A Formal Analysis," in *Essays in Positive Economics*. Chicago 1953.
6. G. FROMM AND P. TAUBMAN, *Policy Simulations With An Econometric Model*. Washington 1968.
7. F. MODIGLIANI AND R. E. BRUMBERG, "Utility Analysis and the Consumption Function," in K. K. Kurihara, ed., *Post-Keynesian Economics*, New Brunswick 1954. Reprinted in H. R. Williams and J. D. Huffnagle, *Macroeconomic Theory: Selected Readings*. New York 1969.
8. D. PATINKIN, *Money, Interest, and Prices*, 2nd ed., New York 1965.
9. William A. SALANT, "Taxes, Income Determination, and the Balanced Budget Theorem," *Rev. Econ. Statist.*, May 1957, *39*, 152–61, reprinted in R. A. Gordon and L. R. Klein, eds., *Readings in Business Cycles*, Homewood 1965.
10. *Federal Reserve Bulletin*, "Consumption and Saving Patterns Since Mid-1968." Washington, June 1969.
11. *Monthly Review*, Federal Reserve Bank of New York, August 1969.

Coinsurance and the Welfare Economics of Medical Care

By MICHAEL CREW*

Recent contributions to the solution of the problem of optimal insurance coverage have formulated rules which an individual might use in determining his optimal quantity of liability insurance.[1] Competition in the "servicing" of liability claims has generally been assumed.[2] The purpose of this note is to demonstrate an apparent paradox. Where monopoly or some restriction of competition exists in the servicing of liability claims, coinsurance[3] may lead to a Pareto optimal situation. Using the assumption of competition in the supply of medical care it is possible to show that, because of the presence of the phenomenon known as "moral hazard,"[4] an optimal situation will not be attained under schemes of insurance or coinsurance.

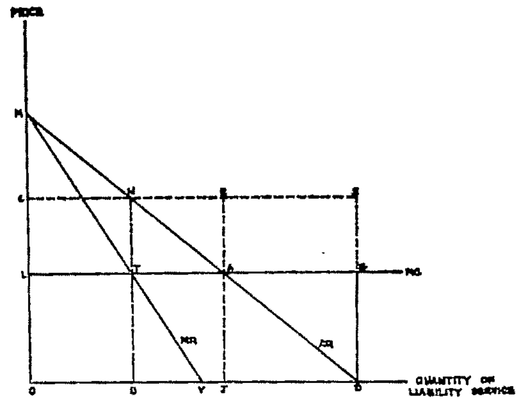


FIGURE 1

The discussion which follows will, like the earlier literature, relate specifically to the problem of medical care insurance. However, other very common insurance services face the problem of moral hazard, for example, automobile collision repairs, and malingering under schemes of compensation for loss of work.

The problem of optimal resource allocation under insurance exists because of the phenomenon of moral hazard. The problem and the solution attained through coinsurance are illustrated in Figure 1. The demand and marginal revenue curves for a medical service are shown as MD and MV respectively. If medical care were supplied at marginal cost of OL, an optimum output would be attained at OJ. The effect of complete insurance would be that insured persons are able to obtain medical care at zero marginal cost. They would, therefore, consume OD, and the efficiency loss of ADG would result. Had the demand curve been a straight line

* Rutherford College, University of Kent, Canterbury. The author was visiting assistant professor of economics at Carnegie-Mellon University. He would like to thank K. J. Arrow, O. A. Davis, L. B. Lave, M. I. Kamien, and L. A. Rapping for their helpful comments on an earlier draft. He alone is responsible for errors. Support for this research was provided by a grant from Resources for the Future, Inc.

[1] Vernon L. Smith [8] and Jan Mossin [6] were concerned with the general problems of how the individual should determine his optimal purchase of insurance. As such, they provide extension of consumer and utility theory. They both have application to the problem of medical care insurance with which this paper is concerned. Mark V. Pauly [7] and Kenneth J. Arrow [1] were concerned solely with the problem of medical care insurance.

[2] For purposes of Smith's model [8] it would not affect his basic predictions although the individual would obviously purchase different quantities under monopolistic supply of liability services than he would under competition.

[3] Coinsurance means that an individual agrees to bear a certain proportion of all insured expenses himself. A deductible [6, pp. 561–3] [7, p. 536], which involves the individual in paying the first x dollars of every allowable claim, makes no contribution to optimality as transactions cost is ignored. Perhaps the deductible's main contribution to optimality occurs when transactions costs are important. They prevent a lot of small claims with high transactions costs.

[4] Pauly has given an explanation of the origin and meaning of this term. "It has been recognized in the insurance literature that medical insurance, by lowering the marginal cost of care to the individual, may increase usage. . . . Insurance writers have tended very strongly to look upon this phenomenon (of demanding more at a zero price than at a positive one) as a moral or ethical problem . . ." [7, p. 535].

of zero elasticity going through J, complete insurance would still have led to an optimum. In the latter situation, no moral hazard would be present. Thus, the extent of the moral hazard can be determined by the elasticity of the demand curve. The greater the moral hazard, the greater the absolute values of elasticity and efficiency loss. The presence of moral hazard, therefore, means that neither insurance nor coinsurance can lead to an optimum output. All coinsurance can do is lead to some output between OJ and OD. Where medical care is supplied at marginal cost, it seems that no schemes of insurance or coinsurance will attain an optimum.

It is possible to show that coinsurance can lead to an optimal output if medical care is not supplied at marginal cost. It is assumed that a powerful monopoly exists in the supply of medical care, so powerful that it can charge well in excess of marginal cost for its services. From the diagram, it can be seen that the monopolist equates marginal cost and marginal revenue and charges OC, that is CL in excess of marginal cost. If individuals coinsure to the extent of OL—that is, the service now costs them OL—it is clear that the optimum output of OJ will be attained. In this situation no insurance with output OB and complete insurance with output OD have a kind of symmetry. They are both at opposite sides of the optimum. Complete insurance is preferred by the monopolist as he can collect a further rent payment of $NSGT$ when compared with what he gets when there is no insurance. Coinsurance enables an optimum to be attained because there is no efficiency loss as the area NRA is not an efficiency loss but a rent payment to the monopolist.

Thus, the situation which results with coinsurance is one similar to the case of a two-part tariff which allows the consumer to buy all he wants at marginal cost for a fixed charge. It is different from the traditional two-part tariff in that the fixed charge is paid to the insurance company while the unit charge only is paid by the individual to the physicians. In addition, the upper limit of the premium, the fixed charge of a two-part tariff, is not just the consumer's surplus extracted via the Hicksian price compensating variation but the sum of the consumer's surplus and the gain from insurance.

The power of the monopoly, and therefore the rent payment, has for illustrative purpose in the diagram been shown as very large indeed. The analysis would, of course, apply to any restriction of entry and competition which resulted in a rent payment. It is not possible at this stage to say to what extent this analysis is applicable to the medical care market in the United States. It does seem to be fairly clear, however, that the restrictive practices of the medical profession do raise prices and result in a rent payment to the profession.[5]

Some conclusions follow. First, the popularity of schemes of coinsurance and the very expensive nature of complete insurance which are common in the United States are in line with the predictions of this model and illustrate the market's way of approaching an optimum. Second, and related to this, it is obvious that where the demand for a particular service is completely inelastic, complete insurance still leads to an optimum. Thus the rather more extensive coverage provided for accidental injury by some insurance schemes is possibly optimal and in line with the prediction of this model. Thirdly, the optimum is a kind which from a distribution point of view is highly inequitable. It involves the transfer of income to the rich, the medical profession, from the sick, a very large subset of which are the relatively poor and the poor. Finally,[6] it should be stressed that the effect of monopoly in the medical care market will affect insurance premiums and it is possible that some individuals who would take on insurance or coinsurance if pricing were at marginal cost might not do so. Thus, a welfare loss would still exist. However, if it is assumed that the demand for insurance is very inelastic, the optimum output is still

[5] This issue has been examined by Friedman and Kuznets [3], H. G. Lewis [5], Fein [2], and Hansen [4]. No definite conclusions emerge partly through inadequacies in data.

[6] I am indebted to K. J. Arrow for pointing this out to me.

approached and a rental transfer remains. Therefore, the paradoxical nature of this result that monopoly and coinsurance can lead to an optimum output while competition and coinsurance or insurance cannot, should not be allowed to obscure the question of distribution which may be more important.

References

1. K. J. Arrow, "Uncertainty and the Welfare Economics of Medical Care," *Amer. Econ. Rev.*, Dec. 1963, *53*, 941–73.
2. R. Fein, *The Doctor Shortage: An Economic Diagnosis*. Washington 1967.
3. M. Friedman and S. Kuznets, *Income from Independent Professional Practice*. New York 1945.
4. W. L. Hansen, *Shortages of Investment in Health Manpower*. Ann Arbor 1964.
5. H. G. Lewis, *Unionism and Relative Wages in the United States*. Chicago 1963.
6. J. Mossin, "Aspects of Rational Insurance Purchasing," *J. Polit. Econ.*, July/Aug. 1968, *76*, 553–68.
7. M. V. Pauly, "The Economics of Moral Hazard," *Amer. Econ. Rev.*, June 1968, *58*, 531–7.
8. V. L. Smith, "Optimal Insurance Coverage," *J. Polit. Econ.*, Jan./Feb. 1968, *76*, 68–77.

On the Social Rate of Discount: Comment

By ALAN NICHOLS*

In his recent article in this *Review* [1], W. J. Baumol errs in his handling of social opportunity costs. The error is also found in a more subtle, because less explicit, form in some recent writings of E. J. Mishan [2] [3], and it should presumably be corrected before it becomes a new doctrine.

On the assumptions of a 50 percent corporate income tax and that all resources are transferred to government from corporations, Baumol concludes that where r is the public's time preference rate for consumption, $2r$ will be the opportunity cost of transfer, since he treats the corporate income tax as being shifted. This is unobjectionable. But he argues further that whether resources are transferred by borrowing or by taxation makes no difference. The reason he believes the treatment is the same in either case arises from his view that it makes no difference whether the resources are drawn to the public sector from private investment or from private consumption. All that matters is that "...this transfer of resources must take place *through the agency of the corporation*" [1, p. 792; italics in original]. He is arguing "...only that the outlays of these firms [on raw materials, here steel] would have brought them a rate of return of *2r as a result of the consumers' marginal valuation of these commodities*" [1, p. 792; italics in original].

If we take the view that the last dollar spent on every type of resource must yield the same return, and we suppose that the last dollar of capital investment yields $2r$ rather than r, this will imply a commensurately higher return on the last dollar's purchase of other inputs as well. An alternative statement of the matter would come from regarding capital as amorphous purchasing power and thus seeing the value of the marginal product per dollar of expenditure on one input as being, with efficient allocation, the same as the value of the marginal product per dollar of expenditure on any other input. Evidently it is on some such grounds as these that Baumol maintains that, in effect, the transfer of one resource from the private sector to the public sector is just like the transfer of any other resource. Again such a view by itself would seem to be unobjectionable. But it would fail to support his conclusion that it does not matter whether resources are transferred to government from private investment or from consumption.

All resources used in the corporate sector may in some sense such as that just discussed yield the same "rate of return," but this is a different matter from the total magnitude of those resources. The fundamental difference between attracting resources from consumption, as would largely be the case with taxation, and attracting resources from saving, as would largely be the case with government security issue, is that a larger quantity of resources would be passed to the future if we use the taxation route. Thus if the question is the shift of resources to future generations, which is what one should think we are discussing when examining the "social rate of discount," it does make a difference which route is chosen. To say, as Baumol says, that "...the costs of taxation versus borrowing are considerations relevant to the choice of stabilization policy. They should not determine whether or not a specific project is undertaken" [1, p. 793] misses the point altogether, and in two ways at that. The less important is that presumably he believes that someone would argue this broad question along the lines that a specific project X should be adopted if, say, it is to be financed by borrowing but not if it is to be financed by taxation. Such things may be important enough for local school districts, but they hardly should affect national policy makers in setting standards for benefit-cost analysis in the context of spending large sums on a variety of projects, especially since no one could say whether any, all, or

* Professor of economics, Central Michigan University.

none of the program was financed by taxes or borrowing, when the Federal government is using both to finance its general budget. The more important point is that "costs of taxation versus costs of borrowing" really does reduce to an intergenerational transfer question, even in the context of stabilization policy. In turn, to bring in such matters as stabilization policy or, say, inflation as a means for capital transfer, is simply to obscure the issue in view.

It may be argued that there is a second fundamental difference between the two methods since one involves loans which are voluntary transfers, while the other involves compulsion, namely, taxes. Presumably the public has more taste for the former than the latter. I believe this differentiation is incorrect. The loan method of finance involves less compulsion now, but later it will require higher taxation to permit debt servicing. Thus the issue is not that between voluntarism and compulsion, but rather between compulsion now and compulsion later. It may be noted, further, that this point is valid even if the debt is never paid off but simply refunded into the indefinite future. The point in question can be seen most simply by supposing that debt issuance is in the form of perpetuities. In this case it is immediately apparent that the present value of future taxes must be calculated at the same rate as the current yield (if we treat the new debt as sold at par), and this means that the current borrowing will necessarily be matched by future taxes of identical present value.

One may argue, and perhaps correctly, as do Baumol and Gordon Tullock, that there is little case for making the future richer at the expense of the present and that, indeed, perhaps the reverse has the better of it. In any event, acceptance of Baumol's underlying view would take from us a "degree of freedom" that is in fact present, for example borrowing rather than taxing to help finance the solution of current social and economic problems. This would, of course, be the reverse of the case of subsidizing the future by financing government capital projects through taxation. A neutral policy would involve financing government capital projects through borrowing and government current account spending through taxation. There are thus more possible combinations than Baumol's analysis tells us. The financing of the spending decision as well as the nature of the spending decision has alternative present-future implications.

A problem seems to remain in that r in this analysis appears to be the borrowing rate: $2r$ is the rate applicable to tax-raised funds. Baumol, as noted, sees these two rates as, respectively, the time preference and the opportunity cost rates. He opts for $2r$ but on grounds "...that are hardly convincing even to me..." [1, p. 801]. The problem of the two rates from the present perspective would be that r would apply to transfers from the future to the present, borrowing, while $2r$ would apply to the reverse transfer, taxing. Fortunately, however, this problem exists only in Baumol's analysis, not in the logic of the process of resource transfer. He is correct in assigning $2r$ to transfers through taxation but incorrect in assigning only r to transfers via borrowing. If it is true that r is the subjective time preference of those who lend in the market, it still does not follow that the government should cost the lending at r; rather the lending should be costed at its opportunity productivity, $2r$, i.e. at what the loans would produce if they were made to the private market. If the opportunity cost rate were not correct, it would pay us to execute large transfers from private to public investment on the basis that every dollar of investment would earn half in the public sector of what it earned in the private sector. This is of course an absurdity. A better case could be made, should we want to make the future richer, for transfering from private consumption. Here, as Baumol says, the rate is $2r$. Surely it cannot be less if we transfer from private investment. The logic of the matter may be put in still a different way. If we suppose that, in principle, interest charges are not always, into the indefinite future, to be borne by further debt, we must suppose them to be ultimately borne by taxes.

Baumol's assumption that the time preference rate is lower than the social opportunity cost rate is the same as saying the public wants more capital formation than it is getting, i.e. that it wants to enrich the future. This assumption, however, can in no sense authorize the government to use the time

preference rate for costing government projects. There is not only the opportunity productivity of direct funds transferred from the private sector of $2r$, there is also the opportunity productivity of making virtually limitless transfers to the private sector that will pay at the rate $2r$, through the simple expedient of debt retirement.

If we relax Baumol's implicit assumption that production is confined to the corporate sector, it might be argued that there would be other opportunity productivity rates that could be used instead of, or in addition to $2r$. Thus partnerships and proprietorships would be pushed to an equilibrium at r rather than $2r$, because no corporate taxes apply. Do-it-yourself activities, in turn, would be pushed to a rate below r, because no income taxes at all apply. From the point of view of efficient allocation, we then have overinvestment in do-it-yourself and in partnerships and proprietorships vis-à-vis corporations. But, does this have a bearing on the discount rate government should use? The answer is no. The opportunity cost with respect to government activity remains $2r$, the corporate sector rate. To see this, one need only note that output could be increased in the private sector by taxing low rate of return activities, and using the proceeds of the taxes to finance high rate of return activities, to wit, those earning $2r$. In such a setting it is obvious that government should not cost its projects at anything less than could be brought about in the private sector itself by the expedient of tax adjustments alone. Or, to look at the same matter in a different way, government could eliminate corporate taxes and treat the reduction in revenue as an investment earning $2r$. So long as such an option is open to government, it makes no sense for government to cost its projects at a lower rate.

There remains the argument that the opportunity cost of a project should be the earnings rate on the funds in the use to which they would actually have been put rather than the earnings rate in their highest rate of return use, i.e., say, r instead of $2r$ in the case of funds transferred from the noncorporate sector. This would dichotomize cost on the basis of source of financing rather than leaving it on the sole basis of next best use, as has been here proposed. Putting aside the question of how a particular source of finance might be assigned to a particular project out of a general budget, there is no doubt that, in the model under review, it would be better to raise funds by taxes in the noncorporate sector; indeed, this is simply the corollary of the position that we could cut corporate taxes and regard the reduced revenue as a capital investment earning $2r$. We could replace the funds by taxing the noncorporate sector and be ahead by the magnitude r per dollar. But this is only a round-about way of saying what the model says at the outset, the corporation income tax is less efficient than other taxes. Now it is easy to confuse two things; 1) how should we cost government spending? 2) how should we tax? Regarding the second, we know the answer at once from the model. Regarding the first, presumably we should spend the limited government revenues in ways which earn the highest rates of return. This last, however, is equivalent to saying that we should use opportunity cost in the sense of "next best use;" but given that there is a corporate income tax, that next best use will earn at least $2r$. Perhaps the corporate income tax should be cut or eliminated; that is one thing. How the government should cost its spending, given that there is a corporate income tax, is something else.

Two conclusions emerge from this note. First, it is not true that it is a matter without consequence whether funds transfers occur through borrowing or through taxing. The different routes have different distributional consequences among generations. Second, the correct rate to use in Baumol's framework is categorically the opportunity rate, $2r$.

References

1. W. J. Baumol, "On the Social Rate of Discount," *Amer. Econ. Rev.*, Sept. 1968, *58*, 788–802.
2. E. J. Mishan, "How to Make a Burden of the Public Debt," *J. Polit. Econ.*, Dec. 1963, *71*, 529–42.
3. ———, "The Burden of the Public Debt: A Rejoinder," *J. Polit. Econ.*, Oct. 1964, *72*, 491–95.

On the Social Rate of Discount: Comment

By ESTELLE JAMES*

In William Baumol's recent provocative article, the corporate income tax is shown to create a divergence between the marginal rate of return to private investment, on the one hand, and the subjective time preference or government borrowing rate, on the other [3]. Baumol claims that the former constitutes the opportunity cost of capital and hence the optimal social discount rate for evaluating public projects, whether the resources used have been diverted from private investment or consumption [3, p. 792]. The following analysis suggests somewhat contrasting conclusions:

1. If we take the tax as given, but assume that government investment is perfectly substitutable for private and will get us back to the pre-tax optimum, the nondistorted interest rate, rather than either of the distorted rates, signals the equilibrium marginal time preference and should be used for discounting. Associated with this rate and the implied consumption-savings ratio are a set of nondistorted prices differing from the distorted prices; the former comprise the correct marginal valuations of goods and factors in calculating the returns to various investment projects.

2. If government investment only partially removes the distortion, marginal rates of substitution still give us relative values of different goods and of consumption in different periods. In other words, the subjective time preference which results after all expenditures should be used for discounting, regardless of the source of resources for public investment. If a government project displaces a private one, society gains when the present discounted value of the former is higher, even though its rate of return is lower.

3. If social risk is zero, the existence of private risk does not alter the appropriate discount rate but may influence the nature of projects in the government's investment portfolio.

4. If public and private goods are nonsubstitutable, a government project should be undertaken if it has a higher present value (but not necessarily a higher rate of return) than the consumption-plus-private-investment bundle from which it draws its resources. The marginal subjective time preference remains the correct discount rate.

I

Assume, as Baumol does, a full employment, risk and externality free economy in which all output is consumed or invested and all income is consumed or saved, the relative proportions varying with the interest rate. Society faces a given array of independent multiperiod investment projects, $I'I'$ in Figure 1, which may be ordered according to marginal rates of return and undertaken by equity-financed private firms or government; the nature of the producing agent does not alter preferences or productivity.[1] The savings schedule, SS, gives the supply of funds available for investment, after netting out interhousehold loans in a perfect capital market. Initially there is no government spending or taxation.

* The author, associate professor of economics at the State University of New York at Stony Brook, wishes to thank Egon Neuberger and Herman and Lois Stekler for their helpful comments.

[1] By "independent" is meant that the adoption of one project does not affect the rate of return to any other project. Thus, mutually exclusive alternatives, including variable factor proportions, are ruled out (see fn. 5). By "multi-period" is meant that we are dealing with the present plus more than one future period. This condition is important for Section II. By "marginal" is meant that each project, while fixed in its capital requirements, is a small part of and incremental to total investment. The following discussion abstracts from distributional considerations; to make unambiguous statements about social welfare we assume that when efficiency rises everyone's utility is increased and vice versa when efficiency falls. This comment deals only with the impact of the corporate income tax and risk, and ignores the influence of externalities and intergenerational effects on the optimal social discount rate.

The competitive equilibrium, shown in Figure 1, implies an interest rate of r' and savings-investment of L'. This is also a welfare optimum; social returns coincide with private, all projects with a yield greater than r' have been undertaken, and all income associated with a subjective time preference less than r' has been saved.

A 50 percent profits tax is suddenly imposed on firms and fully distributed to consumers in the form of transfer payments. The private rate of return is now only half the social, the new investment schedule falling to II. The savings schedule may also shift, but for simplicity we assume it to be constant, i.e., independent of the income distribution. Equilibrium savings and investment are reduced to L, with the marginal social rate of return equalling $2r$, the private return, government borrowing rate, and subjective time preference equalling r, as in Baumol's model.

As a consequence of this drop in the savings-consumption division of income, relative prices, and hence the ordering of projects along the investment schedule, will change; the internal rates of return to different projects depend on the external interest rate. Demand for consumer goods and factor inputs thereto, including capital, will rise, and conversely in the producer goods sector. The new prices correctly inform us only of those marginal valuations associated with the new split-level interest rate.

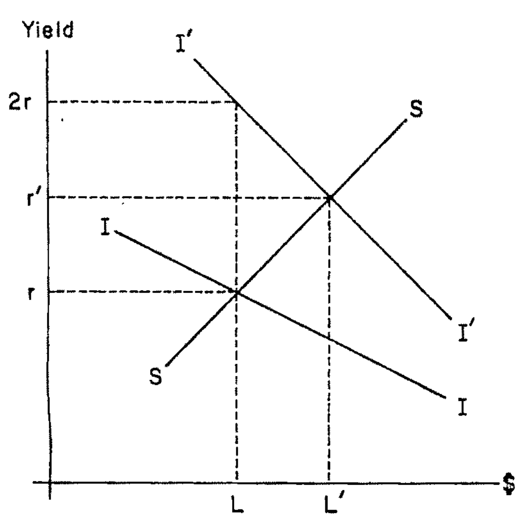


FIGURE 1

How does the new level of welfare compare with the old? We know that (some) people would prefer to save more if they could receive a yield greater than r; at r' they could have saved L but chose to save L'. We also know that projects do indeed exist which would yield society more than r. Optimal intertemporal resource allocation requires that L' be saved and invested, although private firms will no longer get us there because of the tax wedge. The government now undertakes to remove this market failure. The simplest remedy would be to subsidize (or remove the tax on the returns to) private investment, but this is politically infeasible, so we turn to the alternative, public investment, and seek the appropriate discount rate.

Notice that we now have a choice among three social discount rates, not two. And neither r, as in current government practices, nor $2r$, as in Baumol, will get us to optimality. Only r'—the nondistorted interest rate which equates marginal social return to marginal subjective time preference—will accomplish that. Should the government finance its investments by borrowing on the market, r' will reappear as the new equilibrium interest rate. Alternatively, additional savings may be induced through fiscal or regulatory measures (with the government careful not to inefficiently "force" savings upon those with a subjective time preference greater than r'.) Then, r' may be operationally difficult to locate but continues to identify the investment bundle which optimizes welfare.

Only LL' need be invested by the government if all resources are drawn from current consumption and aggregate private investment is unchanged. This is, in fact, difficult to achieve; even consumer goods production requires investment, which must be earning at least a $2r$ return, but may now be displaced by government projects earning less than $2r$. How can such behavior improve the situation, Baumol asks.[3]

[3] Baumol also objects to use of the subjective discount rate on grounds that this rate is substantially influenced by, and constantly adjusting to, a variable

The answer is twofold. First, as noted above, the artificially reduced level of savings and private investment has changed relative prices, costs, and rates of return in different industries. Marginal projects in the consumer goods sector will no longer be earning $2r$ when the government restores consumption and investment to their initial optimal levels.

Secondly, let us suppose that government investment, directly or indirectly, has indeed caused private rejection of a project whose nondistorted yield exceeds $2r$. This will happen, for example, if the government finances its expenditures by borrowing on the market and pushes up the interest rate. Then, the government must adopt the rejected project and invest more than LL' to return the economy to an optimal equilibrium. Under Baumol's rule, paradoxically, a further reduction in private investment would dictate a still higher social discount rate and therefore a lower level of government investment. We have found that additional government investment *can* improve the situation and the appropriate discount rate remains at r', irrespective of the initial source of funds or the ultimate source of resources.

II

Let us continue all the conditions of the preceding section, but assume, more realistically, that the government may not completely remove the distortion and return us to optimality. Decisions are made on each venture separately, and, in the extreme, acceptance of a government project means rejection of a private project with equal capital requirements and a higher rate of return. Does Baumol's choice of $2r$ as the correct discount rate apply here?

It has been carefully demonstrated, although often forgotten in the investment planning literature, that a higher rate of return on a project does not necessarily indicate a higher present discounted value.[3] Ranking by rate of return gives us the entire bundle of investments which should be undertaken at the equilibrium interest rate. However, if only some of these will be accepted, the choice should be based on their relative present values, which depend on the external discount rate, not the internal rate of return. The alternative having a smaller rate of return, because its revenues accrue in the more distant future, may have a higher present value if the discount rate is low; the lower the discount rate the greater the likelihood that these two criteria will diverge.

Thus we are brought back to the very question we started with: What is the correct social discount rate? The problem of determining relative values is a familiar one in economic theory, and generally we solve it by asking at what rates, at the margin, are people willing to trade off one product, or dated consumption, for another. Clearly the latter is given by the marginal subjective time preference, which, before government investment, is r. Discounting at this rate, some projects with a return exceeding $2r$ may be inferior to an alternative yielding less than $2r$ and indeed less than r'.

Society's time preference will rise as the total of public plus private investment increases, ultimately reaching r' if all the conditions of Section I are met. The correct discount rate may therefore be unknown in advance and wrong decisions may be made if each project is evaluated separately at the initial rather than the final discount rate.[4]

Does this analysis imply that the source of resources is irrelevant in public investment planning? Hardly. We may choose a govern-

monetary policy, [3, p. 801]. It should be noted that his discounting choice, the marginal yield to private capital, is a fixed multiple of the subjective time preference and hence equally influenced by monetary policy. This paper abstracts from complications arising in the money market.

[3] See for example, Alchian [1], Bailey [2], Hirshleifer [5], and Baumol [4, pp. 440–45]. The dichotomy holds in a multiperiod planning model where projects differ in their time distribution of net revenues. This paper does not consider additional complexities which arise when the market interest rate varies through time.

[4] Lumpiness in investment projects may in this context lead to index number and convergence problems. See, for example, Bailey, [2]. A parallel statement can be made about the need to use new equilibrium product and factor prices.

ment project which displaces private investment, but would be better off still if we had, at least partially, displaced consumption. The government project may have a higher present value than one private venture but lower than another, even if the two private projects have the same rates of return (but different time distributions of benefits). Therefore, the source of resources for government investment influences how much, if at all, we are improving matters, and hence the desirability of undertaking the project, but it does not alter the appropriate discount rate.[5]

III

Suppose now that risk is present, in varying degrees for different firms, and stockholders demand corresponding risk premiums. Thus, relatively risky firms, which must pay a larger premium, will have a higher marginal capital cost and marginal expected return than risk-free firms. Social risk, however, is zero, whether the project is undertaken by the government or by private firms.[6] Hence this situation is nonoptimal even without a tax, for investment in all firms should be carried to the point where marginal expected returns equals the pure (risk-free) marginal time preference.

The analysis now follows familiar lines. Government can get us to the optimum, but only by discounting at the pure interest rate, rather than following Baumol's advice to include the private risk premium. The government portfolio is then likely to contain a relatively large share of investments judged risky by the private market.

If the government does not fully restore optimality, and displaces private investment, we must compare the expected present values of the relevant public and private projects, again using pure subjective time preference as the appropriate discount rate. This means that the decision is less likely to favor a government project drawing resources from high risk-taking firms.

IV

With government and firms facing the same production possibilities and society indifferent to the nature of the producing agent, we have assumed away the problem, central to Baumol's thesis, of balance between the public and private sectors. To an extent, in the real world, the public and private sectors do indeed produce close substitutes, particularly in fields such as health and education.

Consider now the opposite case: government and firms have completely independent production possibilities, involving basically different, noncompetitive goods and services. As before, imposition of a profits tax introduces a wedge between the subjective time preference and marginal social return on private investment. Since people derive different utilities from collective and private goods, government production can never remove this tax-induced distortion. Thus, the argument developed in Section I is inapplicable, but that in Section II is directly relevant.

Contrary to Baumol, additional government investment is now desirable, providing it has no (direct or indirect) negative effects on private investment. This would, admittedly, produce a higher public/private capital ratio than would prevail in an optimal world. But, by not undertaking the extra investment when its yield exceeds the new lower subjective time preference, the government would only be making matters worse, not better, requiring people to consume (privately and/or collectively) a larger share of income than they desire.

[5] As in the "burden of the tax" literature, we are here concerned with the ultimate source of the resources, rather than the initial source of funds. Incidentally, once mutually exclusive alternatives are introduced, supplementary government investment cannot generally bring us back to the optimum. So long as firms continue to discount at a rate higher than r' due to the tax, they will end up with socially nonoptimal factor proportions and product mix, e.g. relatively labor-intensive techniques and short-lived capital. The existence of private risk, discussed in Section III, has the same distorting effect. This in turn may alter the "second best" total of investment and social discount rate.

[6] I am here ignoring, as Baumol does, the disutility incurred by the risk-averting stockholder, if the project is undertaken by a private firm. To society as a whole the risk is always less—in this case zero—because of the law of large numbers.

More generally, government projects will (e.g. through capital market repercussions) displace some private projects—the trade-off with which Baumol is concerned. We must then compare the present values of the public investment versus consumption-plus-private investment alternatives. In the polar case of fixed total investment, the present value, (discounted at the subjective time preference) of the foregone private project becomes the opportunity cost of the public project, which should consequently be undertaken only if its net benefits are positive.

V

In summary, we have found that:

1. Government investment may displace a combination of private investment and consumption; the ultimate source of resources affects the net gain (or loss) to society, but not the appropriate discount rate.

2. In choosing between public and private projects, decisions based on present value may not correspond to decisions based on rate of return; the former is generally correct.

3. Prices of goods and factors and orderings of projects will shift as a result of interest rate changes brought about by a large government investment program; the new marginal valuations should be used in calculating present values of incremental public and private projects.

It follows that some collective projects should be undertaken with a rate of return lower than that of the marginal private project. On the other hand, some collective projects with a rate of return higher than the equilibrium borrowing rate should not be undertaken. The lower the present value, the more suspect the project, although a rigorous decision cannot be reached unless the precise alternatives are identified, operationally hardly a simple rule. In making these difficult choices, the new marginal subjective time preference remains the correct discounting rate if, as in traditional theory, individual valuations are to count and determine the efficient (intertemporal) allocation of resources.

References

1. A. A. Alchian, "The Rate of Interest, Fisher's Rate of Return Over Cost, and Keynes' Internal Rate of Return," *Amer. Econ. Rev.*, Dec. 1955, *45*, 938–43.
2. M. J. Bailey, "Formal Criteria for Investment Decisions," *J. Polit. Econ.*, Oct. 1959, *48*, 476–88.
3. W. J. Baumol, "On the Social Rate of Discount," *Amer. Econ. Rev.*, Sept. 1968, *58*, 788–802.
4. ———, *Economic Theory and Operations Analysis*, 2nd ed., Englewood Cliffs 1965.
5. J. Hirshleifer, "On the Theory of Optimal Investment Decision," *J. Polit. Econ.*, Aug. 1958, *66*, 329–52.

On the Social Rate of Discount: Comment

By CARL LANDAUER*

William Baumol's article "On the Social Rate of Discount" [1] seems convincing to me in its main argument, but I believe that some of his thoughts may usefully be expanded. He has supplied a convenient point of departure for such expansion when he wrote: "National pride leads many of us to want a promising future for our country."[1] S. A. Marglin has very lucidly expressed this idea in a more general form, when he explained the "schizophrenic answer" to the problem of the social rate of discount and stated that the "Economic Man and the citizen are for all intents and purposes different individuals."[2] Why this is true in regard to time preference is easy to see: The subjective time preference of the individual in private affairs is strongly influenced by the thought of the brevity and uncertainty of the individual life; when judging as a citizen, the individual assumes that the community, in which he takes an interest, will be in existence when he is already mouldering in his grave.

It seems worthwhile to consider one factor which in historically important instances has increased the difference between private and public time preference in the minds of at least some members of the community: A great historical mission which the individual attributes to his community. This, of course, played a major role when communist governments imposed enormous sacrifices on the population for the purpose of rapid industrialization and related objectives. We may condemn a system which enabled a few hundred thousand communist functionaries to impose their low time preference on a population of far more than a hundred millions (the case is that of Marglin's "autocratic answer"), but our political objections do not excuse us from explaining the operation of the sense of national mission where it exists. Moreover, a sense of mission, requiring heavy present sacrifice for the sake of the future, may also exist under democratic government.

If this sense of mission is strong enough, it may sustain the rationality of large present sacrifices for the sake of a distant future in spite of Tullock's objection that " . . .the next generation is . . .going to be wealthier than we are . . ."[3] and that therefore to enrich them further would be, in Baumol's words, a " . . .Robin Hood activity stood on its head."[4] The Tullock argument is in any event open to several objections, setting relatively narrow limits to its applicability. First, there are the irreversibilities mentioned by Baumol: the impossibility of restoring the

* Berkeley, California.

[1] Baumol [1, p. 799].

[2] Marglin [5, p. 98]; see also Landauer [4, pp 21–22]. Marglin feels uncomfortable with the "schizophrenic answer", but, I believe, without good reason. In the difference between the private time preference and that of the citizen, he sees the same inconsistency as between the decision of a man to vote for an ordinance against ticket fixing and his attempt to bribe an officer when stopped for speeding. But the two cases are radically different. The motorist is guilty of a moral inconsistency which may be a little mitigated but is not removed by Marglin's "interdependence answer" (that the motorist would not bribe if nobody bribed.) Since the motorist's support for the anti-bribing ordinance is presumably at least in part based on moral considerations, the moral inconsistency becomes a logical inconsistency. The act which he commits is identical in kind with the act which the ordinance was intended to suppress—the same kind of decision is at stake. This is not true of the time preference problem. If I buy a house now rather than twenty years from now when I shall have saved enough to pay cash, and therefore have to borrow at 8 percent interest, and simultaneously vote for public construction of a dam which will yield a return of only 4 percent, these are two different and essentially unrelated decisions, one of which legitimately influenced by my desire to enjoy the house for a greater part of my limited lifespan, the other just as legitimately influenced by the belief that the community will exist longer than myself. There would be an inconsistency, to be sure, if I first voted for the dam and then cheated on the taxes necessary for its financing, but this possible dishonesty is no essential part of the time preference situation.

[3] Tullock [6, p. 334]. As an old "Austrian," I cannot refrain from reminding the reader that the Tullock argument bears a close relationship to the "first reason" for time preference given by Böhm-Bawerk. See Böhm-Bawerk [3, pp. 440 ff, Engl. version 249 ff].

[4] Baumol [1, p. 800].

fertility of the soil or the beauty of the land once they have been destroyed. The living generation has to preserve these at its own expense if they are to be preserved at all. Destruction of the environment, an inevitable result of unrestrained use of resources for the increase of material output, is part of what Adolf Berle has called "disproduct".[5] Since a portion of it is not preventable and may grow more than proportionately with output of goods and services, there may even be some doubt whether, in a true welfare sense, future generations will really be better off. But, assuming that the growth of material output still leaves an increasing surplus over the "disproduct", there are great uncertainties about future needs: Will the population, for instance, not exceed the optimum? What new goals of perhaps great importance will appear on the horizon of future generations? We have no way of knowing the amount of resources which these needs may demand. For these reasons alone I would add some question marks to Baumol's (tentative and qualified) conclusion that " . . .in our economy, by and large, the future can be left to take care of itself."[6]

It should not be taken for granted that as citizens we have the same time preference in all cases. Some public investments have a close relationship to our private fortunes; whenever this relationship constitutes our principal stake in that investment, it seems that the factors keeping our private time preference high must also have a decisive influence here. My main stake in an improvement of the local fire department is the protection of my own house, an interest that will end with my death, at least if my heirs will not live in the same community and I know they will sell the house. I may still hate the idea that my neighbors may be killed by fire when I am already dead, but I may also feel that I can leave it to them to take the necessary protective measures in the future. On the other hand, there are public investments which have a much looser relationship to our our personal lives, with the consequence that the latter's brevity and uncertainty can have no great enhancing effect on our time preference. My interest in wilderness conservation may be largely independent of my own chance to visit national parks; my patriotic interest in the power, independence, or prestige of the United States may have little to do with any personal satisfaction, such as hearing admiring words from foreigners, and will therefore not be much influenced by my realization of the brevity of my life and will not be rapidly eroded by the passing of time. A convinced communist, although he would wish to see communism established as soon as possible, will still be greatly interested in the prospect of Communist rule a hundred years from now, if it cannot be created earlier. Perhaps the most extreme case would be that of a religious person who wants the divine will to be done in history: Because the deity is eternal, it may not matter greatly to the faithful whether the realization will be achieved 10 years from now or 500 years from now, at least not in regard to the sacrifices he is willing to make. It seems to me that all ideological purposes, in the broadest sense of the word, keep their weight with relatively little reduction through the prospect of time.

If this is correct, it is one more argument—in addition to those adduced by Baumol—why the search for a unitary, optimal rate of social discount is a wild goose chase.

[5] See Berle [2].

[6] Baumol [1, p. 801].

References

1. W. J. Baumol, "On the Social Rate of Discount", *Amer. Econ. Rev.* Sept. 1968 *58*, 788–802.
2. A. A. Berle, "What GNP Does Not Tell Us", *Saturday Review*, Aug. 31, 1968.
3. E. von Böhm-Bawerk, *Positive Theory des Kapitals*, Innsbruck 1909, Engl. Trans. by William Smart, New York 1891
4. C. Landauer, *Theory of National Economic Planning*, 2nd ed, Berkeley and Los Angeles, 1947
5. S. A. Marglin, "The Social Rate of Discount and the Optimal Rate of Investment", *Quart. J. Econ.*, Feb. 1963, *77*, 95–112
6. G. Tullock, "The Social Rate of Discount and the Optimal Rate of Investment", *Quart. J. Econ.*, May 1967, *78*, 331–36.

On the Social Rate of Discount: Comment

By DAVID D. RAMSEY*

William Baumol has written a very provocative article on the social rate of discount [1]. There is no doubt that the specification of the "appropriate" discount rate for evaluating governmental projects has perplexed economists for quite some time and will probably continue to do so. Baumol's major contribution to this problem is his formulation which, unfortunately, he did not fully exploit. Using his formulation, I will show that his conclusions are merely special cases of the general case; in addition the following conclusions emerge: (1) the social rate of discount is a weighted average of observable pre-tax market rates of return and is ***not*** itself directly observable. The "weights" depend on how individuals in the private sector react to the transferral of resources. (2) Both risks and time preferences are given their appropriate consideration in this "weighting" process and neither need to be explicitly considered again. Hence, the indeterminacy or the inconsistency that Baumol found disappears, and the question about what is the appropriate rate of discount becomes an empirical one. Section I contains an analysis of the social rate of discount when there are no risks in the economy and time preferences are ignored. Risks are introduced in Section II and time preferences in Section III.

I. *The Riskless Tax Model*

I start by assuming the conditions of Baumol's basic model with one modification: the addition of a noncorporate sector.[1] I assume that all markets are perfectly competitive. Let X, Y, and Z be the homogeneous outputs of the noncorporate sector, the corporate sector, and the public sector respectively, and let r be the rate of return on noncorporate capital and government securities where the income earned on these securities is nontaxable. Let r_g be the gross rate of return on corporate capital and t be the average and marginal corporate income tax rate. Resources will be allocated between the corporate and noncorporate sector so that the after tax rates of return will be equalized,[2] i.e., $r_g = [1/(1-t)]r$.

Under what conditions will a transfer of resources from the private sector to the public sector be beneficial, neglecting distributional considerations? This, of course, is a familiar problem of public finance where Z is not optimally produced within the private sector so that its evaluation must occur within a nonmarket framework employing some kind of cost-benefit analysis. This problem can be reformulated by simply asking: what rate of return must the production of Z yield to compensate for the amounts of the private goods given up?

Clearly, the answer to the above questions depends on what private goods are given up and how they are evaluated by consumers. It always seems to be implied that a transfer of resources from the private to the public sector comes from the private sector's highest or best alternative. This is the unstated assumption that permits Baumol to state that the return on Z, r_s, must equal or exceed $[1/(1-t)]r$ for the proposed transfer of re-

* The author is assistant professor of economics at the University of Missouri-Columbia. I wish to acknowledge the Business and Public Administration Research Center for financial support and thank Paul E. Smith, James Moore, David Kamerschen, and William J. Baumol for helpful comments. All remaining errors are, of course, mine.

[1] His assumptions are: (1) all resources are fully employed; (2) risk and uncertainty are nonexistent; (3) all private goods and services are supplied by corporations: (4) corporations are financed entirely by equity; (5) corporate income is subject to a 50 percent tax rate; (6) there is a unique interest rate, r, at which the government sells bonds; Baumol [1, p. 792].

[2] Note that consumers' time preference rate will be equal to r *if* banks operate in the noncorporate sector so that r represents individual borrowing and lending rates. If the banks are in the corporate sector then borrowers' time preference rate will be equal to r_g and lender's will be equal to r. But more of this later.

sources to be beneficial. Suppose, without inquiring how or why, resources are diverted from the production of X to the production of Z. Then as long as $r_z \geq r$ the proposed transfer is beneficial. This indicates that the source of "financing" Z *is* important and has a crucial influence on the rate of return that the public project must yield. By "source of financing", I mean the amounts of X and Y given up to get the increased Z and not whether taxes, borrowing, or printing money are used to transfer the resources.

Reconsider Baumol's argument in light of the noncorporate sector. The rate of return earned on resources devoted to the production of X is r which reflects consumers' marginal valuation of X. The rate of return which reflects consumers' marginal valuation of Y is r_g or $[1/(1-t)]r$. If the project is financed wholly by reducing the production of X (all changes will be assumed to be small), then as long as consumers' marginal valuation of ΔZ yields a return greater than r, individuals (or society) will benefit. If ΔZ is financed entirely from reductions in the production of Y, then ΔZ must yield a return of at least $[1/(1-t)]r$ or r_g. If the decline in X production finances a percent of ΔZ and decreases in Y production finances the remainder, then Z must yield a return of at least $ar+(1-a)\ [1/(1-t)]r$ for society to benefit.

To state the argument a bit more formally, let MC_x, MC_y, and MC_z be the marginal *resource* cost of X, Y, and Z respectively excluding "cost of capital." Also I_x, I_y, and I_z are the dollar amounts of investment required to produce one unit of X, Y, and Z respectively. P_x and P_y are the prices consumers must pay. Thus, $(P_x-MC_x)/I_x=r$ or $P_x=MC_x+rI_x$, and $(P_y-MC_y)/I_y=r_g$ or $P_y=MC_y+r_gI_y$. In equilibrium

$$(1)\qquad \frac{MC_x + rI_x}{MC_y + r_gI_y(1-t)} \geq \frac{MC_x(1+r)}{MC_y(1+r_g)} = \frac{P_x}{P_y}.$$

Consumers' marginal valuations of a unit of X and Y are given respectively by $P_x=MC_x+rI_x$ and $P_y=MC_y+r_gI_y$. Consumers face the price ratios on the right-hand side of the inequality in (1) but producers-investors are confronted with the price ratios on the left-hand side of the inequality. Given the price ratios they face, producers and consumers are, at the margin, still indifferent between producing and consuming another unit of X at the expense of a unit of Y. If resources are transferred from the production of X and Y to the production of Z, then the value of the lost X and Y production must be equal to, or exceeded by, the value of the increased Z production if society is to benefit. Thus,

$$(2)\qquad P_x|\Delta X| + P_y|\Delta Y| \leq P_z\Delta Z.$$

The left-hand side of (2) is the value of the reduction in private output (the opportunity cost of Z) whereas the right-hand side is the value of the increased production of Z. Society will benefit only if the inequality obtains. Now (2) can be rewritten as

$$(3)\qquad (MC_x+rI_x)|\Delta X| + (MC_y+r_gI_y)|\Delta Y| \leq (MC_z + r_zI_z)\Delta Z.$$

Thus, r_z, the rate of return on Z generated by consumers' marginal valuation of ΔZ, must satisfy the inequality for society to benefit.

Note that

$$(4)\qquad MC_x|\Delta X| + MC_y|\Delta Y| = MC_z\Delta Z$$

and

$$(5)\qquad I_x|\Delta X| + I_y|\Delta Y| = I_z\Delta Z.$$

Equation (4) simply states that the cost or value of noncapital resources to the public sector must equal their private sector cost or value. The same applies to amounts invested in the public and private sector.[3] Thus, using (4) and (5) and dividing (3) by $I_z\Delta Z$ we get

$$(6)\qquad ar + (1-a)r_g \leq r_z$$

where

$$a = \frac{I_x|\Delta X|}{I_z\Delta Z}.$$

[3] Otherwise resources will not voluntarily leave X and Y production to produce Z. If the government conscripts resources then this valuation still reflects actual opportunity costs.

Since

$$r_g = \left(\frac{1}{1 - t}\right) r$$

(6) becomes

$$r_s \geq \left(\frac{1 - at}{1 - t}\right) r. \tag{7}$$

The right-hand side of (7) is the social rate of discount and it clearly depends on the relative decreases of X and Y production. If we consider Baumol's special case where all of the decrease in output comes from Y, a highly unlikely occurence I would imagine, so that $a=0$ then the social rate of discount becomes $[1/(1-t)]r$. As long as r_s exceeds $[1/(1-t)]r$ then the value of the increment of public output exceeds the value of the foregone private output.

To specify the social rate of discount that leads to an optimal division of resources between the public and private sector requires that we know what portion of the increased public output is financed by reductions in X and Y. Thus the appropriate discount rate lies somewhere in the range between r and $[1/(1-t)]r$; just where in this range depends on the proportional decline in investment in X and Y. Note, however, that Baumol is quite right in asserting that the *form* and *manner* in which X and Y are decreased is irrelevant. What *is* important, which Baumol neglected, is the proportion of Z financed by reductions of investment in X or Y. Whether the reduced investment in X and Y is effected by collecting corporate income taxes, printing money, selling bonds, or conscripting resources is immaterial as Baumol [1, pp. 792–3] pointed out.[4] One further important point: in this model there are two observable market rates of interest, r and r_g, but the social rate of discount is *not* one of these prevailing market rates; it is some linear combination of them and it is not directly observable in the market.

II. *The Role of Risk*

The analysis of risk is analogous to the tax case above. Again consider Baumol's basic model except suppose there is no taxation. Let N be the riskless product, R the risky product, r the riskless rate of return, and r' the risky rate of return.[5] Also let MC_N and MC_R be the marginal costs of the riskless and risky products respectively where the cost of capital or the rate of return is excluded. Given a transfer of resources from the private to the public sector, what return must ΔZ yield to compensate for the reduction of R and N? According to Baumol [1, p. 796], r_s must be at least equal to r'.

The value of the reduction in private output is:

$$P_N \mid \Delta N \mid + P_R \mid \Delta R \mid \tag{8}$$

which must be less than the value of the increased public output, i.e.,

$$P_N \mid \Delta N \mid + P_R \mid \Delta R \mid \leq P_Z \Delta Z, \tag{9}$$

or

$$\begin{aligned}&(MC_N + rI_N) \mid \Delta N \mid \\ &+ (MC_R + r'I_R) \mid \Delta R \mid \leq (MC_z + r_s I_z) \Delta Z\end{aligned} \tag{10}$$

But

$$MC_z \Delta Z = MC_R |\Delta R| + MC_N |\Delta N|$$

and

$$I_Z \Delta Z = I_R |\Delta R| + I_N |\Delta N|$$

so (10)

becomes

$$br + (1 - b)r' \leq r_s, \tag{11}$$

where

$$b = \frac{I_N \mid \Delta N \mid}{I_Z \Delta Z}.$$

Thus,

$$r_s \geq r' + b(r - r') \tag{12}$$

will ensure that society benefits from the resource transfer. The social rate of discount is given by $r' + b(r - r')$ and hence lies in the range between r and r' depending on the

[4] If corporate taxes are increased by increasing t, then the social rate of discount must be recalculated.

[5] It is not relevant for my argument to specify just how risk enters the model. All that is necessary is to note that rates of return do differ, at least partly, because risks differ. See Nerlove [6] and Fisher and Hall [2].

value of b. When $b=0$, Baumol's special case arises where all of the reduction in private output comes from R, i.e., R finances the entire increase in the production of Z and the social rate of discount is r'.

When corporate income taxes are present, four market rates are generated: r, the riskless rate of return on government and noncorporate securities; r', the risky rate of return in the noncorporate sector; r_g, the riskless rate of return in the corporate sector also equal to $[1/(1-t)]r$; and finally, r'_g, the risky rate of return in the corporate sector also equal to $[1/(1-t)]r'$. Here again, the social rate of discount is not a directly observable market rate; it is a linear combination of them. This suggests that if public officials can manipulate the decrease in private output, the social rate of discount can be kept low and many seemingly low return public projects *can* be beneficial.

III. *Time Preferences*

Baumol argues that in the taxless risk case the time preference rate (for both individuals and society) is r, but the social discount rate is r'. The social discount rate cannot simultaneously be r and r'. The reason for the divergence, Baumol [1, p. 798] argued, is that there are two different optimization problems. The r' rate leads to an optimal division of resources between the public and private sector whereas r yields the optimal consumption pattern over time. Hence, one unique social rate of discount cannot simultaneously lead to optimality in both situations. It will be shown below that the time preference rate need not receive any more attention than any other prevailing rate of return. The reason for this is that the consumer's rate of time preference is reflected in observable market rates of return and the above procedure for determining the social rate of discount already takes this into account.

Reconsider the riskless tax case and suppose that only individuals make personal loans; hence income earned from making personal loans is not subject to the corporate income tax. The rate of return on personal loans is therefore r and, assuming that there are no transactions costs, this is the rate that borrowers will have to pay. The time preference rates of both borrowers and lenders at the margin is r. This appears to be just like "Baumol's dilemma." From above, when time preferences were ignored the social rate of discount was

$$\left(\frac{1-at}{1-t}\right)r.$$

But this is not the case because now (accounting for time preferences)

$$\begin{aligned}(13)\quad & r\,|\Delta L| + (MC_x + rI_x)\,|\Delta X| \\ & + (MC_y + r_g I_y)\,|\Delta Y| \le (MC_z + rI_z)\Delta Z\end{aligned}$$

where $r|\Delta L|$ represents borrowers' foregone value of current consumption of X or Y, $|\Delta L|$ is the absolute dollar change in loans. Again, it must be true that

$$(14)\quad MC_x\,|\Delta X| + MC_y\,|\Delta Y| = MC_z\Delta Z,$$

and

$$(15)\quad |\Delta L| + I_x\,|\Delta X| + I_y\,|\Delta Y| = I_z\Delta Z.$$

Hence (13) becomes

$$(16)\quad a_1 r + a_2 r + (1 - a_1 - a_2) r_g \le r_z,$$

where

$$a_1 = \frac{|\Delta L|}{I_z\Delta Z} \quad \text{and} \quad a_2 = \frac{I_x\,|\Delta X|}{I_z\Delta Z}\,.$$

Therefore

$$(17)\qquad r_z \ge \frac{1-(a_1+a_2)t}{(1-t)}\,r$$

The social rate of discount is given by the right-hand side of (17). The social rate of discount as computed above thus reflects the time preference rates of individuals. Moreover, the time preference rate affects the social discount rate in exactly the same way as the rate of return on X. Thus, the social rate of discount is a weighted average of gross or pre-tax rates of return which automatically reflects and gives the appropriate weight to time preferences. This, of course, follows from Baumol's ingenious insight about the necessity for only considering con-

sumers' marginal valuation of the decline in private output *as exhibited by pre-tax rates of return*.

"Baumol's dilemma" may now be resolved as being contradictory special cases. To get a social discount rate of $[1/(1\text{-}t)]r$, (a_1+a_2) must be zero, i.e., all of the decline in current consumption must come from nonborrowers' consumption of Y. To get a social discount rate of r, (a_1+a_2) must be one, but (a_1+a_2) cannot simultaneously be one and zero; in the above formulation it is not.

If banks make loans and are located in the corporate sector and subject to taxation, then a wedge is driven between the time preference rates of borrowers and lenders whose time preference rates are r_g and r, respectively. This is analogous to the riskless tax case where the wedge is between consumer and producer prices; here the "wedge" is between the borrower and lender time preference rates. Reformulating the problem we get

$$(18)\quad r_g|\Delta L| + P_x|\Delta X| + P_y|\Delta Y| \leq P_z\Delta Z$$

which reduces by (14) and (15) to

$$(19)\quad r_g|\Delta L| + rI_x|\Delta X| + r_gI_y|\Delta Y| \leq r_sI_z\Delta Z.$$

Thus,

$$(20)\quad r_s \geq \frac{1 - a_3t}{1 - t}r$$

where

$$a_3 = \frac{I_y|\Delta X|}{I_z\Delta Z}.$$

Now $r_g\ |\Delta L|$ represents borrowers' marginal valuation of their decreased amounts of current X and Y consumption. It matters little to investors or lenders whether their investment in X, Y, or L decreases since, after taxes, their return is the same.

Notice that time preference could have been handled in exactly the same way in a taxless risk model where the risky venture is personal loans; the same results obtain. Risk can also be introduced in the above tax model with the result that for society to benefit

$$(21)\quad r \leq r_s \leq \left(\frac{1}{1 - t}\right)r'$$

Denote the social discount rate by r_s, it then becomes

$$(22)\quad r_s = \frac{r' + a_2(r - r') - (a_1r' + a_2r)t}{1 - t}$$

where

$$a_1 = \frac{|\Delta L|}{I_z\Delta Z} \quad \text{and} \quad a_2 = \frac{I_y|\Delta X|}{I_z\Delta Z}$$

and it is assumed that personal loans involve the same amount of risk as does the production of Y.

IV. *Summary and Conclusions*

There are two basic conclusions: (1) the social discount rate is a linear combination or weighted average of observable market rates of return and hence is *not* a directly observable market rate. The weights depend on the private sector's reaction to a transfer of resources from market-oriented production to nonmarket-oriented production. (2) Time preferences and risks are accounted for in the "weighting" procedure and, hence, need not receive explicit direct consideration again. To summarize, I have shown that the social rate of discount is given by

$$(23)\quad r_s = \sum_{i=1}^{q} a_ir_i,$$

where r_i is the market observable pre-tax rate of return (including risk) in the i^{th} investment venture and where a_i represents the proportion of the publicly directed level of output financed by private sector reductions in the i^{th} output including loans. Of course,

$$\sum_{i=1}^{q} a_i = 1.[6]$$

[6] If some private outputs actually increase so that a set of private output reductions finances both increases in Z and other private inputs then it is possible for r_s to be greater than $1/(1-t)\ r'$ where r' is the riskiest cor-

Thus, risks, time preferences, differential taxes, and the effects of monopoly are taken into account by this procedure and consumers' marginal valuation of the reduction in private sector output is correctly measured.[7]

Baumol concludes his paper by indicating his preference for a social rate of discount at the higher end of the range of interest rates. This, of course, implies that the resource transferral comes mainly from relatively higher risk, higher taxed, private ventures. This may be the case, but it is primarily an empirical question although there are some other implications which can now be stated. The nature of Z probably influences r_s, i.e., the a_i are not unique and invariant with respect to the *kind* of public project being considered. Public projects that draw resources from the low risk, low taxed areas of the private sector will yield a lower r_s than if the opposite case holds. A diversion of resources from the agricultural sector will probably yield a lower r_s than if the resources were diverted from the corporate sector. Also, different methods of financing the resource transferral may yield different values for the a_i. An increase in the corporate income tax will yield a vector of a_i and an increase in the personal income tax will no doubt yield a different vector of a_i.

Determining the social rate of discount along the lines suggested here is somewhat analogous to determining the incidence of a tax. A tax usually causes price and output changes and the purpose of tax incidence analysis is to identify the resulting price and output changes. Hence, it may be fruitful to apply the theory of tax incidence to the determination of the social rate of discount.[8] Also, the recent literature on the firms' cost of capital appears to have a direct bearing on the approach used here.[9]

porate investment or for r_s to be less than r, the noncorporate sector's riskless rate of return. However, given usual convexity assumptions, this appears unlikely.

[7] Note that the same formal results are obtained under monopoly as were obtained in the riskless tax case. Also, prices, marginal costs, and rates of return may be changed by the resource transfer because of the nature of production and utility functions; this introduces an additional problem of choosing between before-project and after-project prices to evaluate the reduction in private output.

References

1. W. J. Baumol, "On the Social Rate of Discount," *Amer. Econ. Rev.*, Sept. 1968, *58*, 788–802.
2. I. N. Fisher and G. R. Hall, "Risk and Corporate Rates of Return," *Quart. J. Econ.*, Feb. 1969, *83*, 79–92.
3. A. C. Harberger, "The Incidence of the Corporation Income Tax," *J. Polit. Econ.*, June 1962, *70*, 215–40.
4. P. N. Mieszkewski, "On the Theory of Tax Incidence," *J. Polit. Econ.*, June 1967, *75*, 250–62.
5. M. H. Miller and F. Modigliani, "Some Estimates of the Cost of Capital to the Electric Utility Industry, 1954–57," *Amer. Econ. Rev.*, June 1966, *56*, 333–91.
6. M. Nerlove, "Factors Affecting Differences among Rates of Return on Investments in Individual Common Stocks," *Rev. Econ. Statist.*, Aug. 1968, *50*, 312–31.

[8] See Harberger [3] and Mieszkowski [4].

[9] See Miller and Modigliani [5].

On the Social Rate of Discount: Comment

By DAN USHER*

William Baumol raises the intriguing problem of choosing a rate of interest for the public sector when risk or tax interposes a spread between the subjective rate of time preference and the technical rate of opportunity cost between present and future consumption in the private sector [1]. A number of writers have suggested that the rate of time preference is appropriate for discounting projects in the public sector. Baumol first goes to the other extreme by proposing that government projects ought to be discounted at a rate representing opportunity cost in the private sector. Later in the article, he takes the more moderate position that both subjective time preference and opportunity cost are relevant. Baumol states that,

> ... there is an unavoidable indeterminacy in the choice of that rate. The figure which is optimal from the point of view of the allocation of resources between the private and public sectors is necessarily higher than that which accords with the public's subjective time preference. As a result, neither the higher nor the lower figure that has been proposed can, by itself, satisfy the requirements for an optimal allocation of resources, and we find ourselves forced to hunt for a solution in the dark jungles of the second best ... [1, p. 789]

There is some question as to what is meant by indeterminacy in the first sentence quoted above. Baumol may be alluding to the divergence of views among economists on the appropriate rate of discount for government projects. The choice of the rate cannot be literally indeterminate if the correct rate can be hunted for in the dark jungles of the second best. Even when the rate of time preference differs from the rate of opportunity cost in the private sector, an optimal rate of discount for the public sector is implied by the rules for maximizing the welfare of the community in the choice of public and private goods. Using the method developed by Richard Lipsey and Kelvin Lancaster [2], I shall show how to choose the interest rate on government projects and will prove that, under certain quite general assumptions, the appropriate interest rate on government projects lies between the rate of time preference and the rate of opportunity cost between present and future consumption in the private sector.

Imagine an economy in a world with two time periods, 0 and 1, and two types of commodities, private goods that must be produced in the private sector and public goods that must be produced in the public sector.[1] The choice between public and private goods is made by the government in tax and expenditure policy. The government maximizes a community welfare function

$$U(C_{0p}, C_{0g}, C_{1p}, C_{1g}) \tag{1}$$

subject to a production possibility constraint

$$T(C_{0p}, C_{0g}, C_{1p}, C_{1g}) = 0 \tag{2}$$

where C_{0p} is current consumption of private goods,
C_{0g} is current consumption of public goods,
C_{1p} is future consumption of private goods,
and C_{1g} is future consumption of public goods.

[1] The distinction between public and private goods is essential in this analysis. Without the distinction, or something like it, a spread between rates of interest in the public and private sectors would cause welfare to be maximized when all investment is undertaken in the public sector. To justify assuming that all private goods are produced in the private sector, it might be said that the private sector is more efficient at producing private goods or that excessive expansion of public activity would have undesirable social or political consequences. It would be odd to base a case for public enterprise on misallocations arising from the tax system.

* Professor of economics, Queen's University, Kingston, Ontario.

The first order condition is[3]

$$\frac{U_{0p}}{T_{0p}} = \frac{U_{0g}}{T_{0g}} = \frac{U_{1p}}{T_{1p}} = \frac{U_{1g}}{T_{1g}}$$

To ensure that the community welfare function is well-defined, it is assumed that everyone in the economy has the same tastes and assets. The government need choose only the amounts of public consumption, C_{0g} and C_{1g}, and the private sector can be relied upon to choose the correct amounts of private consumption, C_{0p} and C_{1p}. It is convenient to think of C_{1p} and C_{1g} not as consumption next year, but as flows of consumption in perpetuity from next year on, so that derivatives like dC_{1p}/dC_{0p} and dC_{1g}/dC_{0g} may be looked upon as rates of interest.

There are twelve prices in this economy, six rates of transformation in production and six rates of substitution in use. Each set of six prices includes two ordinary relative prices, two own-rates of interest, and two cross-rates of interests. Suppose that C_{0p}, C_{0g}, C_{1p} and C_{1g} are chosen. In production, the ordinary prices R_{0t} and R_{1t}, are the relative prices of public and private goods now and in the future. By differentiating equation (2), these are seen to be

$$R_{0t} = \left.\frac{dC_{0p}}{dC_{0g}}\right|_T = \frac{T_{0g}}{T_{0p}} \tag{4}$$

$$R_{1t} = \left.\frac{dC_{1p}}{dC_{1g}}\right|_T = \frac{T_{1g}}{T_{1p}} \tag{5}$$

The own rates of interest, r_{pt} and r_{gt}, are the rates of transformation in production between present and future consumption of private and public goods respectively.

$$r_{pt} = \left.\frac{dC_{1p}}{dC_{0p}}\right|_T = \frac{T_{0p}}{T_{1p}} \tag{6}$$

$$r_{gt} = \left.\frac{dC_{1g}}{dC_{0g}}\right|_T = \frac{T_{0g}}{T_{1p}} \tag{7}$$

[3] The term U_{0p} means $\partial U/\partial C_{0p}$, the terms U_{0g}, U_{1p}, U_{1g}, T_{0p}, T_{0g}, T_{1p} and T_{1g} are defined accordingly; the functions T and U are defined so that these terms are positive.

The cross rates of interest, r_{pgt} and r_{gpt}, are

$$r_{pgt} = \left.\frac{dC_{1g}}{dC_{0p}}\right|_T = \frac{T_{0p}}{T_{1g}} \tag{8}$$

$$r_{gpt} = \left.\frac{dC_{1p}}{dC_{0g}}\right|_T = \frac{T_{0g}}{T_{1p}} \tag{9}$$

The six rates of transformation in use, R_{0u}, R_{1u}, r_{pu}, r_{gu}, r_{gpu}, r_{pgu} are defined with respect to the welfare function, and it follows from the first order conditions, equation (3), that all rates of transformation in use are equal to the corresponding rates of transformation in production. However, the own-rates are not equal; it is easily seen that the own-rates on public consumption and on private consumption are connected by the formula

$$r_{gt} = r_{pt}\frac{R_{0t}}{R_{1t}} \tag{10}$$

Equation (10) indicates that the own-rate of interest on public goods exceeds the own-rate on private goods if public goods are getting cheaper as time goes on.

The reason for drawing attention to the distinction between own-rates and cross-rates and to the fact that these rates are not equal, is that a rate must be chosen to represent Baumol's notion of opportunity cost of investment in the public sector. In particular, the rate must have the property that it would equal the private rate of return to investment in the absence of institutional constraints interposing a gap between the rates of transformation in production and in use between present and future consumption in the private sector.

To simplify the problem and to introduce an institutional constraint on private investment, we make the following assumptions:

(a) The rate of transformation in production between public and private goods is unity in both time periods.

$$R_{0t} = R_{1t} = 1 \tag{11}$$

As this is assumed true for all choices of output, it implies that the form of the production possibility constraint is

$$T(C_{0p} + C_{0g}, C_{1p} + C_{1g}) = 0. \tag{2'}$$

It follows from this assumption that all four rates of interest in production are equal.

$$r_{pt} = r_{gt} = r_{pgt} = r_{gpt} \tag{12}$$

The rates of interest in use will continue to differ when the institutional constraint is introduced.

(b) The production possibility curve is concave in the sense that an increase in the output of any good increases the cost of producing it. In particular,

$$\frac{\partial}{\partial C_{0p}}(r_{pt}) > 0 \quad \text{and} \quad \frac{\partial}{\partial C_{1p}}(r_{pt}) < 0. \tag{13}$$

An increase in C_{0p}, causes an increase in the relative price of C_{0p} and C_{1p}, and an increase in C_{1p} causes a decrease in the relative price of C_{0p} and C_{1p}. Production possibility curves are usually assumed to possess this property. From assumptions (a) and (b) together, it follows that

$$\begin{aligned} \frac{\partial}{\partial C_{0g}}(r_{pt}) &= \frac{\partial}{\partial C_{0p}}(r_{pt}) > 0 \quad \text{and} \\ \frac{\partial}{\partial C_{1g}}(r_{pt}) &= \frac{\partial}{\partial C_{1p}}(r_{pt}) < 0. \end{aligned} \tag{14}$$

(c) The utility function is convex in the sense that an increase in the output of any good lowers its marginal rate of substitution in use with respect to any other good. In particular

$$\frac{\partial}{\partial C_{0p}}(r_{pu}) < 0 \quad \text{and} \quad \frac{\partial}{\partial C_{1p}}(r_{pu}) > 0. \tag{15}$$

An increase in C_{0p}, lowers the relative price of C_{0p} and C_{1p}, and an increase in C_{1p}, increases the relative price of C_{0p} and C_{1p}.

(d) Public consumption now is a closer substitute for private consumption now than it is for private consumption in the future, but private and public consumption now are less than perfect substitutes.

$$\frac{\partial}{\partial C_{0p}}(r_{pu}) < \frac{\partial}{\partial C_{0g}}(r_{pu}) < 0. \tag{16}$$

An increase in C_{0g} causes C_{1p} to become relatively more valuable than C_{0p}, but the effect

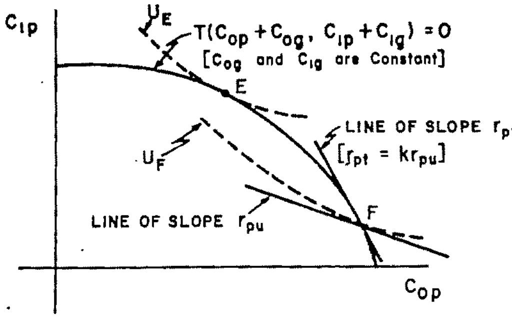


FIGURE 1

of an increase in C_{0g} on the relative valuation of C_{0p} and C_{1p} is less than the effect of an increase in C_{0p}. Similarly, it is assumed that

$$\frac{\partial}{\partial C_{1p}}(r_{pu}) > \frac{\partial}{\partial C_{1g}}(r_{pu}) > 0. \tag{17}$$

(e) Risk or the tax system interposes a spread between the rate of transformation in production and the rate of substitution in use between present and future goods in the private sector.

$$r_{pt} = kr_{pu} \tag{18}$$

where $k > 1$. For example, a corporation income tax of 50 percent causes r_{pt} to be twice the value of r_{pu}; the k corresponding to a corporation income tax of 50 percent is 2. The situation may be illustrated in a two-dimensional diagram, Figure 1. Suppose the government has chosen C_{0g} and C_{1g}. The unbroken curved line, T, is the production possibility curve of C_{0p} and C_{1p}. The broken curved lines, U_E and U_F, are indifference curves. The equilibrium without the constraint is at E where the indifference curve, U_E, is tangent to the production possibility curve. The equilibrium with the constraint is at F where the slope of T is r_{pt} and the slope of the indifference curve U_F, is r_{pu}.

When there is no constraint on private investment, i.e., when $k=1$, the government can maximize U by its choice of C_{0g} and C_{1g} alone because the private sector can be counted on to choose C_{0p} and C_{1p} to equate rates of transformation in production and in use between these variables. When $k>1$, the value of U that the government can obtain

by choosing C_{0g} and C_{1g} is less than the value it could obtain by choosing C_{0g}, C_{1g}, C_{0p} and C_{1p}.

The task of the government is to choose C_{0g} and C_{1g} to maximize

$$(1') \qquad U(C_{0p}, C_{0g}, C_{1p}, C_{1g})$$

subject to

$$(2') \qquad T(C_{0p} + C_{0g}, C_{1p} + C_{1g}) = 0$$

and to

$$(19) \qquad M = kr_{pu} - r_{pt} = 0$$

where, as may be deduced from our assumptions[3],

$$(20) \qquad M_{0p} < M_{0g} < 0$$

and

$$(21) \qquad M_{1p} > M_{1g} > 0$$

Form the Langrangean expression

$$(22) \qquad U - \lambda T - \mu M$$

and differentiate the Lagrangean with respect to the four items of consumption. The new equilibrium conditions are

$$(23) \qquad U_{0p} - \lambda T_{0p} - \mu M_{0p} = 0$$

$$(24) \qquad U_{1p} - \lambda T_{1p} - \mu M_{1p} = 0$$

$$(25) \qquad U_{0g} - \lambda T_{0g} - \mu M_{0g} = 0$$

$$(26) \qquad U_{1g} - \lambda T_{1g} - \mu M_{1g} = 0$$

[3] The term M_{0p} means $\partial/\partial C_{0p}$ (M), and M_{0g}, M_{1p}, and M_{1g} are analogously defined. To prove equation (20), note from equation (19) that

$$M_{0p} = k \frac{\partial}{\partial C_{0p}} (r_{pu}) - \frac{\partial}{\partial C_{0p}} (r_{pt})$$

and

$$M_{0g} = k \frac{\partial}{\partial C_{0g}} (r_{pu}) - \frac{\partial}{\partial C_{0g}} (r_{pt})$$

Assumption (b) states that

$$\frac{\partial}{\partial C_{0p}} (r_{pt}) = \frac{\partial}{\partial C_{0g}} (r_{pt}) < 0$$

and assumption (d) states that

$$\frac{\partial}{\partial C_{0p}} (r_{pu}) < \frac{\partial}{\partial C_{0g}} (r_{pu}) < 0$$

Therefore $M_{0p} < M_{0g} < 0$ as indicated in equation (20). Equation (22) is proved the same way.

As our measure representing Baumol's optimal rate of discount on public investment, we choose, r_{gu}, the own rate of substitution in use between public consumption now and public consumption in the future. Assumption (a) ensures that this rate would equal rates in the private sector if not the for institutional constraint.

We shall now show that optimal rate of interest on public investment is not indeterminate, and that the rate lies between the rates of interest on private consumption in production and in use.

$$(27) \qquad r_{pu} < r_{gu} < r_{pt}$$

To prove that the inequality (27) holds, we first show that $\mu/\lambda > 0$, then we show that $r_{gu} < r_{pt}$, and finally we show that $r_{gu} > r_{pu}$.

From equations (23) and (24), it follows that

$$(28) \qquad r_{pu} = \frac{U_{0p}}{U_{1p}} = \frac{T_{0p}}{T_{1p}} \left[\frac{1 + \frac{\mu}{\lambda} \frac{M_{0p}}{T_{0p}}}{1 + \frac{\mu}{\lambda} \frac{M_{1p}}{T_{1p}}} \right]$$

or

$$(29) \qquad r_{pu} = r_{pt} \left[\frac{1 + \frac{\lambda}{\mu} \frac{M_{0p}}{T_{0p}}}{1 + \frac{\lambda}{\mu} \frac{M_{1p}}{T_{1p}}} \right]$$

By assumption, $r_{pu} < r_{pt}$, $T_{0p} > 0$, $T_{1p} > 0$, $M_{1p} > 0$ and $M_{0p} < 0$. Consequently $\lambda/\mu > 0$, for otherwise equation (29) would be inconsistent.

To prove $r_{gu} < r_{pt}$, note that

$$(30) \qquad r_{gu} = \frac{U_{0g}}{U_{1g}} = \frac{T_{0g}}{T_{1g}} \left[\frac{1 + \frac{\mu}{\lambda} \frac{M_{0g}}{T_{0g}}}{1 + \frac{\mu}{\lambda} \frac{M_{1g}}{T_{1g}}} \right]$$

or, since $T_{0g} = T_{0p}$ and $T_{1g} = T_{1p}$,

$$(31) \qquad r_{gu} = r_{pt} \left[\frac{1 + \frac{\mu}{\lambda} \frac{M_{0g}}{T_{0g}}}{1 + \frac{\mu}{\lambda} \frac{M_{1g}}{T_{1g}}} \right]$$

But $M_{0g}<0$, $M_{1g}>o$ and all the terms T_{0g}, T_{1g} and λ/μ are positive. Consequently $r_{gu} < r_{pt}$.

To prove that $r_{gu}>r_{pu}$, we show that the ratio of these rates is greater than unity.

$$\frac{r_{gu}}{r_{pu}} = \frac{U_{0g}}{U_{1g}} \div \frac{U_{0p}}{U_{1p}}$$

$$(32) \qquad = \frac{T_{0g}}{T_{1g}} \left\{ \frac{1 + \dfrac{\mu}{\lambda}\dfrac{M_{0g}}{T_{0g}}}{1 + \dfrac{\mu}{\lambda}\dfrac{M_{1g}}{T_{1g}}} \right\} \div \frac{T_{0p}}{T_{1g}} \left\{ \frac{1 + \dfrac{\mu}{\lambda}\dfrac{M_{0p}}{T_{0p}}}{1 + \dfrac{\mu}{\lambda}\dfrac{M_{1p}}{T_{1p}}} \right\}$$

By assumption (a), $T_{0g}=T_{0p}$ and $T_{1g}=T_{1p}$, so that

$$(33) \qquad \frac{r_{gu}}{r_{pu}} = \frac{1 + \left(\dfrac{\lambda}{\mu T_{0p}}\right) M_{0g}}{1 + \left(\dfrac{\lambda}{\mu T_{0p}}\right) M_{0p}}$$

$$(34) \qquad \cdot \frac{1 + \left(\dfrac{\lambda}{\mu T_{1p}}\right) M_{1p}}{1 + \left(\dfrac{\lambda}{\mu T_{1p}}\right) M_{1g}}$$

From equations (21) and (22), it follows that both terms on the right-hand side of this expression are greater than 1. This completes the proof that

$$(27) \qquad r_{pu} < r_{gu} < r_{pt}$$

A similar relation holds among r_{pu}, r_{pt} and the rate of transformation between private goods in the present and public goods in the future.

References

1. W. J. Baumol, "On the Social Rate of Discount," *Amer. Econ. Rev.*, Sept. 1968, *58*, 788–802.
2. R. Lipsey and K. Lancaster, "The General Theory of the Second Best," *Rev Econ. Stud.*, Dec. 1956, *24*, 11–32.

On the Social Rate of Discount: Comment on the Comments

By WILLIAM J. BAUMOL*

It is gratifying to be the author of a paper that has led to comments of the quality printed here. Though I differ with a number of the viewpoints expressed, I think on most of them the reader is best left to judge for himself. There is only one issue on which I want to offer a few remarks. In my paper, I addressed myself to the analytic difficulty posed by the possibility that the subjective time discount rate is well below the opportunity cost rate of resources drawn from the corporate sector, and remarked that the choice of social discount rate becomes a matter of the theory of the second best. Dan Usher takes up the challenge and tackles the difficult task of an explicit second best analysis, and argues with its help that " . . . under certain quite general assumptions, the appropriate interest rate on government projects lies between the rate of time preference and the rate of opportunity cost between present and future consumption in the private sector."

David Ramsey, confining himself to the opportunity cost approach, in a world in which the returns to resources differ among sectors of the economy, concludes that " . . . the social discount rate is a weighted average of observable pre-tax market rates of return." It is, in effect, an average of the opportunity costs associated with each of the sources of funds utilized by a government project, each opportunity cost weighted by the proportion of these funds drawn from the corresponding source. This suggests that " . . . public projects which draw resources from low risk, low taxed areas will yield a lower [discount rate] than if the resources were diverted from [other sectors]."

* The author is a member of the faculty at Princeton University.

Alan Nichols, on the other hand, takes a diametrically opposite view. He argues that " . . . the correct [social discount] rate . . . is categorically the [highest] of the pertinent opportunity costs." The argument is, apparently, that if resources are being employed in two private uses R and S where they yield respective returns of r and s percent, then (if $r > s$) funds should not flow to a government project unless it too offers at least r percent, even if the funds are derived from S, for in that case it would be better to transfer resources from S to R than from S to the government project.

Obviously neither of the authors is wrong. Nichols is a perfectionist who will have no truck with second best solutions, while Ramsey will accept any transfers to the public sector provided they constitute improvements. In this imperfect world, my own inclinations lie with Ramsey—I would be unhappy to see opportunities for a better use of resources passed up in an unwillingness to compromise with ideals [1].

Finally, I cannot resist thanking Estelle James for calling to our attention and analyzing an aspect of the problem I had overlooked completely—the case where governmental and private outputs are substitutes. In that case the issue, essentially, is not which collection of services should be produced, but who should provide them. And here, clearly, the relevant considerations are not the same as those I discussed.

REFERENCES

1. W. J. BAUMOL, "On the Discount Rate for Public Projects," in *The Analysis and Evaluation of Public Expenditures: The PPB System*, U.S. Congress, Joint Economic Committee, 91st Cong., 1st sess., Vol. 1, Washington 1969, pp. 489–503.

New Thoughts About Inferior Goods

By H. H. LIEBHAFSKY*

When we combine assumptions that the utility function is additive and that one commodity is an inferior good (defined as one for which purchases decrease as income increases), we produce a case in which there are n-1 inferior goods, each of which has diminishing marginal utility, and one normal commodity (defined as one for which purchases increase as money income increases), which has increasing marginal utility. This result is of considerable general interest. It provides an analytical method of evaluating the results of empirical studies of demand based upon additive utility functions written for blocks of commodities [2] [4] [7]. Unless such empirical studies produce a result in which all income elasticities are positive, they must produce a result in which there are n-1 negative income elasticities and one positive income elasticity.

In addition to the general demonstration mentioned above, this paper also presents what is apparently the first published specific utility function, together with its associated demand functions to illustrate the case of a commodity with a negatively sloping income consumption curve. This specific (additive) utility function can be subjected to a monotonic transformation by squaring it; such a transformation leaves the demand functions unchanged and, in our case, will produce an illustration of the case of an inferior good based on an assumption of dependence of the marginal utilities. I turn first to the general demonstration that the combined assumptions: (1) that the utility function is additive, and (2) that one good is inferior, imply that there are n-1 inferior goods (all with diminishing marginal utility) and one normal commodity (with increasing marginal utility).

Assume the existence of a consumer with a utility function of the form:

$$(1) \quad U = f_1(x_1) + f_2(x_2) + \cdots + f_n(x_n),$$

where U represents total utility and $x_1, x_2, \ldots, x_n$ represent quantities of the goods purchased.

Further, let I represent the consumer's money income and $p_1, p_2, \ldots, p_n$ be the prices of the goods bought. The budget equation can then be written:

$$(2) \qquad p_1x_1 + \sum_{i=2}^{n} p_ix_i = I.$$

The marginal utilities of the goods are designated by

$$(3) \quad u_1 = \frac{\partial U}{\partial x_1}, \ u_i = \frac{\partial U}{\partial x_i} \cdot \quad (i = 2, 3, \ldots, n)$$

The second partial derivatives of the utility function are defined as:

$$(4) \qquad u_{11} = \frac{\partial^2 U}{\partial x_1^2}; \qquad u_{ii} = \frac{\partial^2 U}{\partial x_i^2} \cdot$$

The first order maximization condition is written in the usual well-known way as:

$$(5) \qquad \frac{u_1}{p_1} = \frac{u_i}{p_i} = m,$$

where m is the marginal utility of money.

The conditions of stability deduced by E. Slutsky are applicable to this case [6]. According to Slutsky, it is necessary in our case for the budget to be stable that

$$(6) \qquad d^2U = u_{11}dx_1^2 + u_{22}dx_2^2 + \cdots + u_{nn}dx_n^2 < 0.$$

Slutsky has asserted that $d^2U<0$ if u_{11} and all the u_{ii} are negative and that in such a case $\partial m/\partial I$ will also be negative. He has also asserted that the budget will be stable if one good only has increasing marginal utility, provided that also, $\partial m/\partial I>0$; but that if two or more of the marginal utilities increase, the budget "can never be stable." These stability conditions are of much importance

* The University of Texas at Austin.

$$
(8)\qquad
\begin{aligned}
u_{11}\frac{\partial x_1}{\partial I} + 0\,\frac{\partial x_2}{\partial I} + \cdots + 0\,\frac{\partial x_n}{\partial I} &= p_1\frac{\partial m}{\partial I}\\
0\,\frac{\partial x_1}{\partial I} + u_{22}\frac{\partial x_2}{\partial I} + \cdots + 0\,\frac{\partial x_n}{\partial I} &= p_2\frac{\partial m}{\partial I}\\
\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots&\cdots\\
0\,\frac{\partial x_1}{\partial I} + 0\,\frac{\partial x_2}{\partial I} + \cdots + u_{nn}\frac{\partial x_n}{\partial I} &= p_n\frac{\partial m}{\partial I}\,.
\end{aligned}
$$

in the analysis which follows and should be carefully noted now.

Now differentiate the budget equation with respect to money income (I) to produce:

$$
(7)\qquad p_1\frac{\partial x_1}{\partial I} + \sum_{i=2}^{n} p_i\frac{\partial x_i}{\partial I} = 1.
$$

As is well-known, not all of the terms on the left side can be negative; either $\partial x_1/\partial I$ or at least one of the $\partial x_i/\partial I$ must be positive. Thus, since an inferior good is defined as one such that the relevant $\partial x/\partial I<0$ and a normal commodity is one such that $\partial x/\partial I>0$, equation (7) leads to the conclusion that every case must at least include one normal good, $\partial x/\partial I>0$, because the right-hand side of that equation consists of the positive number, one.

Differentiation of equation system (5) under the assumption of independence ($u_{12}=u_{21}=0$) produces equation system (8) from which we may derive:

$$
(9a)\qquad \frac{\partial x_1}{\partial I} = \frac{p_1}{u_{11}}\,\frac{\partial m}{\partial I}\,;
$$

$$
(9b)\qquad \frac{\partial x_i}{\partial I} = \frac{p_i}{u_{ii}}\,\frac{\partial m}{\partial I}\cdot\quad (i = 2, 3, \ldots, n)
$$

Recall now Slutsky's statement of the stability condition: If u_{11} and all the u_{ii} are negative, so is $\partial m/\partial I$, and "the budget is stable." Now, since p_1 and p_i are positive, if u_{11}, all the u_{ii}, and $\partial m/\partial I$ are negative, the right-hand sides of equations (9a) and (9b) will both be positive. Therefore, so will the left-hand sides be, and thus, in such a case, $\partial x_1/\partial I>0$ and $\partial x_i/\partial I>0$. All n commodities will be normal. Since it is well-known that demand curves of normal commodities have negative slopes, it is clear that the combined assumptions of (1) independent marginal utilities and (2) diminishing marginal utilities for all goods are alone sufficient to produce a case involving n goods having negatively sloped demand curves. Since Marshall made these combined assumptions in showing that demand curves would be negatively sloped, his further assumption that the "marginal utility of money is constant" was redundant.

Equations (9a) and (9b) show that in order to produce a case of an inferior good, $\partial x_1/\partial I<0$ or one of the $\partial x_i/\partial I<0$, either u_{11} or one of the u_{ii} must have a sign opposite from that of $\partial m/\partial I$. But Slutsky's stability conditions rule out the possibility that more than one of the goods can have increasing marginal utility. They require in addition that, if one good does have increasing marginal utility, $\partial m/\partial I$ must be positive also. Moreover, it is not possible to assume that all goods have decreasing marginal utilities and that $\partial m/\partial I>0$. Given the stability conditions, one good only can have increasing marginal utility, and, in this case, $\partial m/\partial I>0$ also. Thus, there must also be n–1 goods with decreasing marginal utility. From equation (9) it follows that, under these conditions, the good with increasing marginal utility will be normal ($\partial x/\partial I>0$) and the n–1 goods with decreasing marginal utilities will be inferior ($\partial x/\partial I<0$). Thus, if any one good is inferior, there will always be n–1 inferior goods whose marginal utilities decrease and *one* normal commodity whose marginal utility increases.

I turn now to the specification of a particular utility function to illustrate the general case; the function used will produce a negatively sloping income consumption curve. Let x_1 represent the quantity of one specific use of a consumer's money income, I. Also let x_2 represent an amount of all other possible uses of the consumer's money income. The prices are again represented respectively by p_1 and p_2, and the budget equation reads:

$$(10) \qquad p_1x_1 + p_2x_2 = I.$$

Assume further that the consumer has the following utility (or preference) function:

$$(11) \qquad U = A_1 \log_e X_1 + \frac{x_2^2}{2},$$

where A_1 is a positive constant.

From (11) we have the marginal utility functions as

$$(12) \quad u_1 = \frac{\partial U}{\partial x_1} = \frac{A_1}{x_1}; \quad u_2 = \frac{\partial U}{\partial x_2} = x_2.$$

From (12) we also have:

$$(13) \qquad u_{11} = \frac{\partial^2 U}{\partial x_1^2} = -\frac{A_1}{x^2} < 0;$$

$$u_{22} = \frac{\partial^2 U}{\partial x_2^2} = 1 > 0;$$

$$u_{12} = \frac{\partial^2 U}{\partial x_1 \partial x_2} = 0 = u_{21}.$$

Our utility function is additive and involves independence, since $u_{12} = u_{21} = 0$.

The demand functions can be produced in terms of money income and prices by substituting from (12) into (5) and solving first for either x_1 or x_2. The result can then be substituted into (2). Since I have illustrated this procedure in detail elsewhere [5, p. 114 ff.], I will merely write the final results in equations (14) and (15). From (14) we can obtain:

$$(16) \qquad \frac{\partial x_1}{\partial I} = \left[\frac{1}{2p_1}\right] \cdot \left[1 - \frac{I}{(I^2 - 4A_1p_2^2)^{1/2}}\right] < 0,$$

since $4A_1p_2^2 > 0$.

Thus x_1 has a negatively sloping income demand curve and is an inferior good.

From (15) we see that x_2 is a normal commodity, since:

$$\frac{\partial x_2}{\partial I} = \left[\frac{1}{2p_2}\right] \cdot \left[1 + \frac{I}{(I^2 - 4A_1p_2^2)^{1/2}}\right] > 0.$$

As stipulated by our general analysis, the inferior good has decreasing marginal utility and the normal commodity has increasing marginal utility. Both price demand curves are negatively sloping, as the reader can determine for himself by taking the necessary

---

$$(14) \qquad x_1 = \frac{I}{2p_1} - \sqrt{\frac{I^2}{4p_1^2} - \frac{A_1p_2^2}{p_1^2}} = \frac{I}{2p_1} - \sqrt{\frac{I^2 - 4A_1p_2^2}{4p_1^2}}$$
$$= \frac{I}{2p_1} - \frac{1}{2p_1}\sqrt{I^2 - 4A_1p_2^2} = \frac{1}{2p_1}\left(I - \sqrt{I^2 - 4A_1p_2^2}\right).$$

$$(15) \qquad x_2 = \frac{I}{2p_2} + \frac{p_1}{p_2}\sqrt{\frac{I^2}{4p_1^2} - \frac{A_1p_2^2}{p_1^2}} = \frac{I}{2p_2} + \frac{p_1}{p_2}\sqrt{\frac{I^2 - 4A_1p_2^2}{4p_1^2}}$$
$$= \frac{I}{2p_2} + \frac{p_1}{2p_1p_2}\sqrt{I^2 - 4A_1p_2^2} = \frac{1}{2p_2}\left(I + \sqrt{I^2 - 4A_1p_2^2}\right).$$

derivatives. Also, the inferior good, X_1, has unitary price elasticity, because of the assumption of unitary elasticity in its utility function, a result not generally true in the case of inferior goods. The reader who wishes to produce a case involving nonzero cross partial derivatives can do so by squaring our utility function. Such a monotonic transformation leaves the demand functions unchanged, a result well-known.

Our analysis provides an answer to the question of the type of utility function necessary to produce a case of an income consumption curve for a good which is normal at low levels of income but inferior at high levels of income, as depicted in several price theory books [1, p. 622], [3, p. 78], [8, p. 85]. If the utility index is additive, such a case must involve the existence of one commodity with decreasing marginal utility and another whose marginal utility function first decreases, next reaches a minimum, and then increases, as equation (9) discloses. The commodity whose marginal utility constantly decreases will be a normal good as long as the marginal utility of the other commodity decreases and $\partial m/\partial I<0$; it will be independent of income when the marginal utility of the other good is at its minimum point and $\partial m/\partial I=0$; and it will be an inferior good when the marginal utility of the other good increases and $\partial m/\partial I>0$. The commodity whose marginal utility decreases, reaches a minimum, and then increases will be a normal good at all levels of income. The analysis of inferior goods in the case of a utility function with nonzero cross partial derivatives by use of Slutsky's stability conditions (the general case of dependence) provides an interesting possibility for useful further study, and I present it as a challenge to the reader. The monotonic transformation produced by the squaring of equation (11) may be one starting point for such an analysis.

References

1. K. E. Boulding, *Economic Analysis*, Vol. 1, New York 1966.
2. S. M. Goldman and H. Uzawa, "A Note Separability in Demand Analysis," *Econometrica*, July 1964, *32*, 387–98.
3. J. R. Hicks, *Value and Capital*, London 1948.
4. H. S. Houthakker, "Additive Preferences," *Econometrica*, Apr. 1960, *28*, 244–56.
5. H. H. Liebhafsky, *The Nature of Price Theory*, rev. ed., Homewood 1968.
6. E. Slutsky, "On the Theory of the Budget of the Consumer," in G. Stigler and K. Boulding, eds., *Readings in Price Theory*. Homewood 1952.
7. R. H. Strotz, "The Empirical Implications of a Utility Tree," *Econometrica*, Apr. 1957, *25*, 269–80.
8. D. S. Watson, *Price Theory and Its Uses*. New York 1963.

The Contribution of Education to the Quality of Labor: Comment

By EDWARD F. DENISON*

David Schwartzman argues in a recent issue of this *Review* [5] that the estimate given in my *Sources of Economic Growth* [1] for the rise in the quality of labor due to additional education was far too big. His comment appeared nine months after, but was written before, publication of my *Why Growth Rates Differ* [3]; I shall also refer to that book in order to bring the discussion up to date.[1]

I must first correct Schwartzman's summary of my substantive results. He reports [5, p. 508] that in *Sources* "... Denison ascribes over three-fifths of the growth in output per man-hour to increased education." My estimates actually implied 29 percent: .67 percentage points [1, Table 33] out of a growth rate of output per man-hour in the whole economy of 2.34 percent.[2]

In *Why Growth Rates Differ*, my procedures were altered to reduce the effect of changes in days of school attended per school year and, with more data available, to incorporate some refinements; for the 1950–62 period they yielded .49 percentage points as the contribution of education [3, Table 21–1]. If applied in the 1929–57 period, the change in my allowance for school days would cut my estimate of .67 percentage points to about .5, or 21–22 percent of the growth of output per man-hour.

In discussing Schwartzman's specific points, I shall usually refer to *growth rates* of educational quality indexes for civilian labor. These exceed *contributions* of increased education of the labor force to growth rates of national income because 1), civilian labor represents only a fraction of total factor input, and 2), in some activities the change in output is measured by employment or man-hours so a change in labor quality cannot affect measured output. My educational quality index for civilians grew at a rate of .75 in 1950–62 [3, Table 8–5, column 4] while the contribution was .49 percentage points. Thus, any change in the civilian quality index would alter the contribution by about one-third less.

My quality index is constructed in two stages. First, I compute an index based only on changes in the distribution of the labor force by years of education. Second, I allow for increased school attendance per school year. Schwartzman does the same. Growth rates of civilian quality indexes compare as follows:

	Estimate Based on Changes in Years (I)	Allowance for Attendance per Year (II)	Final Quality Index
Denison [3] 1950–62	.52	.23	.75
Schwartzman, 1940–60[a]	.20	.13	.33

[a] Schwartzman's calculations are for 1940–60; he lets them refer to 1929–63 on the assumption that "... the rate of change was uniform over the entire period 1929 to 1963 ..." [5, p. 510].

I shall concentrate on this comparison, but note that Schwartzman's own comparison was between his rate, published as .3, and the 1930–60 projection given in *Sources*, which had a growth rate of .94 [1, Table 9].

I. *Years of School*

Schwartzman's Stage I estimate of .20 for 1940–60 is 62 percent lower than my .52 for

* The author is a senior fellow at The Brookings Institution. The views expressed do not purport to represent the views of the other staff members, officers, or trusteees of The Brookings Institution.

[1] Schwartzman's cooperation in providing additional information about his procedures and offering valuable comments on an earlier draft of this communication are gratefully acknowledged.

[2] This figure can be derived from [1, Table 31], which shows growth rates of 2.93 for real national income, 1.31 for employment, and −.73 for average annual hours.

1950–62. Derivation of such estimates requires weights to combine education groups, distributions of the labor force by years of education for dates to be compared, and a procedure to compute a quality index from this information. Schwartzman's criticisms of my Stage I estimates refer chiefly to my education weights, but we also use different distributions and procedures.

Education Weights

In *Why Growth Rates Differ*, weights were derived from Census data for 1959 money earnings of males who had money earnings in 1959, were in the experienced civilian labor force in April 1960, and were 25 to 64 years of age in April 1960. For males within each 10-year age bracket within the 25–64 range, I computed the mean money earnings of those in each years-of-education group as a percentage of mean earnings of those with 8 years of education [3, Table F-1]. The four sets of percentages were then averaged to obtain average percentages standardized by age. These percentages were not used as weights directly because males of the same age with different amounts of education differ not only in education but also in various other characteristics that affect earnings and are correlated with amount of education. In an effort to approximate earnings differentials associated with differences in amounts of education alone, I reduced the earnings differential between each other group and the 8-year group by two-fifths [3, Table F-2].[3] The procedure used in *Sources* differed only in that the data referred to 1949 money income (not earnings) of all males in the four 25-to-64 age groups in 1950 (not to males in the experienced labor force with money income in the prior year).

Schwartzman's objections that the 1949 data used in *Sources* referred to income rather than earnings and included males 25–64 not in the 1950 labor force are valid in principle. But they do not apply to the 1959 weights developed in *Why Growth Rates Differ* and evaporate in practice, because the two sets of weights are so similar [3, Table F-2, columns 2 and 4] that the choice does not affect quality indexes. (I continued to use the earlier weights.) One probable bias mentioned by Schwartzman remains: the less educated among the 1960 labor force probably had more weeks of unemployment in 1959 than the more educated. Schwartzman's wording may inadvertently suggest that he also criticizes me (erroneously) for not standardizing earnings by age groups, but he tells me that this was not his intent. We agree on the desirability of age standardization.

To facilitate discussion of remaining issues, I show in Table 1 weights computed from four sets of earnings data for males 25 to 64. All are based on the same mechanical procedure of computing earnings at each educational level as a percentage of earnings at the 8-year level in each of the four age groups, averaging these percentages, and reducing differentials from the 8-year level by two-fifths. Column 1 gives the weights derived in *Why Growth Rates Differ* from average annual earnings in all civilian industries as calculated by the Census Bureau. Column 4 gives weights based on average hourly earnings in nonfarm industries as calculated by Victor Fuchs in [4]; Schwartzman uses these earnings. Earnings used in columns 2 and 3 are described subsequently. Below each set of weights I show the growth rate, based on these weights, from October 1948 to March 1965 of a quality index for the male civilian labor force 18 years of age and over.[4] The .52 rate based on my weights (column 1) happens to be the same as my 1950–62 rate for both sexes combined, and the effect on these two rates of changing weights would be the same.[5] The rate shown in the last column is .35; thus the difference in selection of earnings as weights accounts for over half the difference between Schwartz-

[3] The assumption is discussed in [1, pp. 68–70, 79]; [2, pp. 86–100]; and [3, pp. 83–84].

[4] The 1948 and 1965 labor force distributions are given in [3, Table F-9].

[5] A change in weights introduces roughly the same percentage change in a quality index regardless of time period (within recent decades). Since, in this case, the level of the growth rates is the same, the change would also be about the same when measured in percentage points.

TABLE 1—ALTERNATIVE EARNINGS WEIGHTS AND ASSOCIATED EDUCATION QUALITY INDEXES

	Census All-Industry Annual Earnings (1)	Census Nonfarm Annual Earnings (2)	Fuchs Nonfarm Annual Earnings (3)	Fuchs Nonfarm Hourly Earnings (4)
Years of education				
0	71	75	75	78
1–4	77	81	81	85
5–7	90	91	91	93
8	100	100	100	100
9–11	110	108	107	105
12	121	120	117	112
13–15	139	136	133	126
16+	181	176	162	150
Male quality index, 1948–65 (growth rate based on years only)	.52	.49	.43	.35
Same, with substitution in 16+ row of column 1 weights	.52	.50	.49	.44

man's Stage I estimate and mine (.17 out of .32 percentage points).

Differences among growth rates shown in Table 1 stem mainly from differences in earnings differentials between the college-trained and those with only elementary education. It is between these groups that the net shift in the distribution is most pronounced and the difference in weights greatest. As shown by the last row of Table 1, substitution of the 16+ weight from the first column for that in the other columns (without altering the weights of the other education classes) would in itself eliminate over half the differences among the rates.

The earnings Schwartzman uses differ in three ways from those I use: 1. Schwartzman excludes agricultural workers. He argues that their earnings are understated relative to nonfarm earners because income in kind is omitted and because the self employed underreport; and that this exaggerates earnings differentials by education level since agricultural workers have less education than others [1, pp. 509–10]. But there are two large offsets to the biases he notes. First, earnings of farm operators include returns from land and capital as well as from labor. Second, earnings of farm operators, who are mostly males 25–64, include income derived from work performed by unpaid family workers as well as from their own work.[6] Some net bias may be present, but my judgment has been that earnings from labor are lower in agriculture, and, contrary to the view implicit in Schwartzman's comment, earnings differentials in the whole civilian economy are better approximated by data for the whole civilian economy than by data for nonfarm industries. I therefore prefer to use all-industry data.[7] In practice, the choice

[6] Unpaid family workers provide no offset. All but a negligible 28,000 are excluded from the data I have used because they are females, or males under 25 or over 64. The remainder are excluded unless they had money earnings.

[7] Schwartzman refers only to biases in the Census farm earnings data, but some readers might suppose there is a different reason to base earnings differentials on nonfarm data: that labor misallocation and perhaps a preference for farm life depress farm earnings and thus widen observed earnings differentials among education groups in the whole labor force. But this is one of several intercorrelations between education and other income determinants for which my two-fifths downward adjustment of earnings differentials is intended to compensate. The more homogeneous with respect to farm-nonfarm composition, race, region, parents' background, and natural ability is the universe for which earnings differentials are computed, the smaller is the appropriate downward adjustment. If I were to start with a uni-

is only moderately important; the growth rate of a quality index would be reduced 6 percent (.03 percentage points in 1948–65) if nonfarm weights were substituted (Table 1, columns 1 and 2).

2. Twice as important is the source of nonfarm earnings. I have relied on Census Bureau data for mean earnings.[8] Schwartzman used means, computed by Fuchs for his study of region and city size, which yield narrower earnings differentials among education groups.[9] Substitution of the Fuchs means depresses the growth rate of the quality index by an additional 12 percent—.06 percentage points in 1948–65, of which .05 points stems from the college graduate group (Table 1, columns 2 and 3).

The difference derives from the midpoints used to compute means from grouped data. Almost half of all males earning over $15,000 were college graduates, so earnings differentials among education groups are sensitive to choice of midpoints in the top earnings classes. Fuchs used $18,500 as the midpoint for the $15,000 to $24,999 earnings class, the Census Bureau $19,000. More important, for the open ended $25,000 and over class Fuchs used a uniform $40,000 whereas the Census Bureau obtained a midpoint for each age-sex-education group by "interpolation . . . from a Pareto curve fitted to the upper income range"; the higher pitched an earnings distribution, the higher was the midpoint used for the top earnings class. With agriculture included, the Census Bureau means imply that earnings in the $25,000 and over class averaged about $46,000 for all males, $43,000 for males other than college graduates, $49,000 for college graduates, and as much as $58,000 for college graduates 55 to 64 years of age.

In practice, almost the entire difference between the quality indexes in columns 2 and 3 of Table 1 reflects Fuchs' lower mean earnings in the two age-education cells in which earnings are highest: college graduates 45–54 and 55–64 years of age.[10] The Census Bureau procedure is generally regarded as preferable. The alternative is a short cut that almost inevitably results in understatement of earnings differentials.

3. Schwartzman uses hourly instead of annual nonfarm earnings. This substitution lowers the quality index by .08 percentage points in 1948–65 (Table 1, columns 3 and 4), or by 15 percent of the original rate. The reduction reflects, to an unknown extent, incorporation in the Fuchs hourly earnings estimates of an adjustment for weeks worked in 1959 as well as that for weekly hours.

Appropriate treatment of hours of work is the most interesting topic raised by Schwartzman's comment. Average hours worked by the more educated exceed those worked by the less educated. If hours for each education

verse more homogeneous in these respects, such as the nonfarm labor force, I would reduce the downward adjustment of earnings differentials, not the spread in the final education weights. For brevity, the two-fifths adjustment is sometimes called an ability adjustment (Schwartzman does so), but analyses I have made or seen suggest two-fifths is far too big an adjustment to compensate for ability, as measured by such indicators as rank in class or IQ scores, alone.

[8] The indexes before the "two-fifths" adjustment are given in [3, Table F-2, col. 5]. (The "college 4 or more" index was corrected from 220 to 226 in the third printing.) The 0–7 years of education group was subdivided into three groups for the present note.

[9] Most of the data are given in [4, p. 45]. Fuchs kindly provided information to divide his 5–8 years of education group into 5–7 and 8 year groups. I estimated the division of his 0–4 groups into two subgroups.

[10] Census data yield mean earnings of $13,388 for college graduates in the 45–54 age group and $13,212 in the 55–64 group after "farmers and farm managers" and "farm laborers and foremen" are eliminated. Fuchs' corresponding nonfarm means are only $12,107 and $11,316. (Figures for high school graduates at these ages are similar.) Although Census nonfarm means cannot be recalculated with Fuchs' midpoints from published data, the difference in midpoint obviously causes the discrepancy. Substitution of Fuchs' midpoints would reduce the Census all-industry means for male college graduates in the 45–54 age group from $13,313 to $11,901 and in the 55–64 group from $13,089 to $11,484.

Fuchs' nonfarm earnings figures differ from those I calculated from Census data in three additional respects. They are based on the .1 percent sample instead of the 5 percent sample. They exclude members of the experienced labor force who were not at work in the 1960 Census week for which hours were obtained. They exclude from the all-industry aggregate persons in the "agriculture, forestry, and fisheries" industry division rather than "farmers and farm managers" and "farm laborers and foremen." The net effect of these differences on the weights appears to be negligible.

group were stable over time, then average hours for the whole labor force would rise as the education distribution moved up. If one of the effects of additional education is to induce people to work longer hours, then to combine an educational quality index based on annual earnings differentials with an hours index based on average hours not standardized by level of education will, as Schwartzman argues, overstate the rise in labor input. Either the education index or the hours index should be reduced to avoid double counting, the choice depending in principle on where one wishes to classify the effect of education on hours.

The latter was my preference; in *Why Growth Rates Differ* I stated: "Insofar as an increase in the education of the labor force raises hours worked above what they would be otherwise, this is classified as an effect of increased education." However, I did not actually adjust the hours index and therefore admitted that the total increase in the quality of a year's work was "slightly overstated" [3, p. 85].

The reasons I did not adjust the quality index for hours were lack of relevant data and my belief that an appropriate adjustment was small. An outer limit can be put on the size of such an adjustment by using Fuchs' hours data to make a mechanical calculation, paralleling that for education, of the effect of changes in the education distribution on average hours worked; this requires an assumption that exclusion of farming from the hours data does not affect the results. Such a calculation shows that average hours of the male labor force 18 and over would have increased .11 percent a year from 1948 to 1965 if average hours worked at each education level had not changed.[11] Suppose this figure were the same for 1950–62, the same for full-time as for all workers, and the same for full-time females (who receive only a small weight in any case) as for full-time males. This would mean that full-time hours of each sex standardized by education dropped at a rate .11 greater than the unstandardized data I used. With my allowances for a change in output per hour as hours change, the 1950–62 growth rate of my series representing the quality index for a year's work as affected by hours would be reduced by about .07. The contribution to the growth rate made by the change in hours would be reduced by about .05 percentage points (from −.17 to −.22) as would the contribution of total labor input (from 1.12 to 1.07). But .05 is too much and perhaps no better an estimate than .00. Fuchs' hours differentials are too large for such an adjustment of the hours series for three reasons. First, farmers are excluded. Second, differences not only in weekly hours but also in weeks worked during 1959 are reflected [4, p. 37], and changes in weeks worked are already measured in the employment component of labor input. Third, the cross-sectional relationship between education and hours may be due as much to concentration of highly educated workers in occupations in which customary hours (at least as reported) happen to be long as to any tendency for more education to induce longer hours.

Schwartzman prefers to adjust the educational quality index instead of the hours index. To do so, he substitutes differentials in hourly earnings for differentials in annual earnings in the construction of weights while retaining the two-fifths downward adjustment for factors other than education. As noted, this substitution (which also adjusts for the apparent bias due to differences in weeks worked) reduces the 1948–65 growth rate of the educational quality index by .08 percentage points, implying a reduction of .05 in the contribution of changes in years of education. These numbers are too big insofar as hours and weeks worked differentials by educational level in the nonfarm economy exceed those in the whole civilian economy. They are also too big to fit into the structure

[11] Hours were computed by dividing Fuchs's annual earnings by his hourly earnings [4, pp. 40, 45, and unpublished data provided by Fuchs]. The weights (with hours worked by those with 8 years of education equal to 100) obtained by averaging percentages for the four age groups within the 25–64 range are: 0–4 years of education, 92.2; 5–7 years, 96.2; 8 years, 100.0; 9–11 years, 103.5; 12 years, 107.3; 13–15 years, 107.4; 16 or more years, 110.2. These weights were applied to the same October 1948 and March 1965 distributions of the labor force by level of education that were used in constructing education quality indexes in Table 1.

TABLE 2—PERCENTAGE DISTRIBUTIONS OF CIVILIAN EARNINGS BY SEX AND AGE

	Age							
	19 and Under	20–24	25–34	35–44	45–54	55–64	65 and Over	Total
Total civilian earnings equal 100 percent								
Males	1.2	4.1	17.7	23.1	19.6	12.0	2.6	80.3
Females	1.0	1.9	3.6	5.0	4.9	2.8	.5	19.7
Civilian earnings of each sex equal 100 percent								
Males	1.5	5.1	22.1	28.8	24.4	15.0	3.2	100.0
Females	5.3	9.8	18.1	25.2	24.8	14.1	2.7	100.0

of my estimates because their calculation assumes that output (and earnings) of those with long hours would be reduced proportionally if their hours were shorter, whereas my procedures imply a substantial productivity offset to differences in hours.[12]

Time Periods

Of the difference between Schwartzman's Stage I rate for 1940–60 and mine for 1950–62, about .15 percentage points are not due to his use of different earnings as weights. The effect of the difference in time periods cannot be calculated exactly but it may legitimately account for over half of this remaining discrepancy. Application of the weights based on Fuchs' nonfarm hourly earnings to decennial Census data for the experienced civilian labor force (excluding emergency workers in 1940) yields a growth rate, .26, for both sexes combined from 1940 to 1960 that is .09 points lower than the rate (which I take to be equal to the rate for both sexes combined in 1950–62) that is shown in Table 1, column 4. This is so even though this 1940–60 rate is overstated by exclusion of emergency workers in 1940. (Based on my weights, the rate for both sexes is .39 in 1940–60 as against .52 in 1950–62; the reduction is about one-fourth with each set of weights.) However, it should be noted that these 1940–60 rates are computed from labor force distributions that come from a different source than the others, and include 14 to 17 year olds.

Labor Force Distributions

My quality indexes are based on use for all dates of consistent distributions of the civilian labor force 18 years of age and over, by years of education, obtained from the household surveys. Schwartzman presents decennial Census distributions in his Table 2 that are not comparable in 1940 and 1960.[13] However, he informs me that these are not the distributions he used in his calculations; those he did use are even less comparable. For 1940 he added emergency workers (apparently his own estimates) to Census data for the employed and experienced unemployed in all industries; for 1960 he used a distribution, based on the 1 in 1,000 sample, that covered only *employed* workers in *nonfarm* industries. This use of noncomparable data must have biased his indexes upward because both the unemployed and the farm

[12] Schwartzman's application of the same (two-fifths) proportional downward adjustment of differentials in hourly as in annual earnings may introduce a slight bias in the other direction; it is arguable that for consistency the downward adjustment should be a little bigger for hourly earnings.

[13] The title of the table erroneously describes the 1940 data as referring to the experienced labor force and the 1960 data as referring to employed persons. Actually the 1960 distributions are for all experienced workers whereas the 1940 distributions cover only employed persons, exclusive of emergency workers as well as the unemployed. If, as his description on page 510 seems to imply, he had used these data, his statement that "the change in the definition of labor exaggerates the increase" in education would therefore be backwards. Also, in his 1960 columns, but not the 1940 columns, persons with 7 years of education are erroneously shown to have only 5 or 6 years.

labor force have below-average education. The effect on a comparison of our estimates was apparently more than offset by a difference between our computation procedures.

Computation Procedures

Schwartzman criticizes me for not treating age groups separately in the computation of quality indexes. He asserts—on what evidence I do not know—that the "... greatest increase in education has been among young persons." That this, even if true, could have biased my series in *Sources* upward, as he states [5, p. 513], is hardly credible because in that study I completely omitted persons under 25 years of age from my distributions, but the procedural question is nevertheless worth examining.

If necessary information were available, an education quality index for all civilians could best be obtained by 1) computing a separate quality index for each civilian age-sex group, using education weights appropriate for that group, and 2) computing a weighted average of these indexes, using total earnings of each group in a base period as weights.[14] Such age-sex weights are given in Table 2.[15] The procedure an investigator follows for a real situation necessarily depends upon the data he can obtain, his evaluation of their accuracy, and his judgment as to the degree of refinement that is worthwhile given data limitations. Evidently Schwartzman and I differ in our evaluations and judgments; he uses data I regard as worse than useless for the purpose at hand.

In *Why Growth Rates Differ*, I computed only two separate quality indexes, one for each sex. I omitted persons 14 to 17 years of age, whose share of employment far exceeds their trivial earnings weight, from the labor force distributions. The same education weights were used for each sex. These were based only on earnings of males 25 to 64 because I regarded data for other groups as unusable; males 25 to 64 account for over 72 percent of total civilian earnings and over 90 percent of male earnings.[16] Computed earnings for groups below 25 years of age at the Census date surely should not be used for education weights. Changes in individuals' labor force status (including the crucial shift from holding a job, usually part-time, while in school to working after leaving school) and changes in educational attainment between January of 1959 (the year to which earnings refer) and April 1960 (the date at which individuals were classified) make computed earnings differentials among education groups meaningless. Obstacles to an accurate adjustment to bring earnings into temporal conformity with labor force and educational status, let alone to estimation of "ability differentials" for age classes with many members in school, seem insurmountable. In addition, earnings of young workers rise very rapidly with age, and the only available data have far too wide age bands (14–19 and 20–24) to permit age standardization. Rapid labor force turnover and frequency of part time work also distort data for females 25 to 64 and persons over 65 (for whom a too broad age range is an additional problem). Distortion is less than for the young, and special tabulations of Census data might provide a basis for adjustment, but this has not been attempted. Female earnings patterns in Fuchs' raw nonfarm data look odd, but in any case do not suggest that female earnings differentials by level of education in the 25–64 age range are systematically wider or narrower than male differentials.

[14] I assume here that, as in [3], the effect of changes in hours worked and age-sex composition have previously been measured separately.

[15] A six-way distribution (between sexes and among three age groups for each sex) for 1962 was carefully prepared for use in [3, Table 7-6]. The further breakdown shown in Table 2 was less carefully prepared using 1960 data from Fuchs and the Census.

[16] An unweighted average of percentages for the four age groups within the 25–64 range was used, as already noted. Alternatively, the percentages for each age group might have been weighted by its earnings, with the "under 25" group's weight assigned to the adjacent 25–34 group and the "over 65" weight to the 55–64 group. With male weights from Table 2 applied, this procedure yields education weights for males that differ from those given in Table 1, column 1, by more than one point only for those with 16+ years of education (178 as against 181). For females, the weights differ by more than one point only for those with 13 to 15 years (137 as against 139) and 16+ years (176 as against 181). This refinement is scarcely warranted in view of the range of error in the data, including the "two-fifths" adjustment, but in any case would not change quality indexes noticeably.

A separate quality index for each age-sex group could, it is true, be calculated without additional earnings information, by use of education weights for each 10-year male age group within the 25–64 range for persons of each sex in that age group, the 25–34 weights for younger groups, and the 55–64 weights for older groups. These indexes could then be combined by earnings weights such as are given in Table 2.[17] My only objection to this procedure is that the task of estimation becomes formidable if one must develop, for a number of dates, labor force distributions cross-classified not only by sex and by years and (as Stage II of my procedure requires) days of education, but also by age.

Although the Schwartzman procedure, used to compute a 1940–60 quality index, is much more detailed than mine (except for use of one education class less), I believe it is subject to much greater error. He initially computes:

$$\frac{\sum_i n_{it}w_i}{\sum_i n_{i0}w_i}$$

where n_i is the percentage of workers who are in each age-sex-education group i (starting with 14–19 age groups), w_i is average hourly earnings of members of group i in 1959, t is the year 1960, and 0 is the year 1940.[18] This ratio is 1.049, representing a 4.9 percent increase from 1940 to 1960. The percentage reflects deterioration in age-sex composition of the labor force as well as the rise in education. By standardization, Schwartzman estimates the effect of this deterioration at 1.5 percent, adds this to the initial percentage, and obtains 6.4 percent as the change due to shift in educational composition. This is multiplied by .6 with the intent of duplicating my two-fifths downward adjustment of earnings differentials.[19] He thus obtains a Stage I educational quality index that rises about 4 percent in 20 years, representing a growth rate of .20[20] From the comparisons already given, it is apparent that this is less than application of my procedure to the same distributions of the labor force in 1940 and nonfarm employment in 1960 would yield.

My chief objection to Schwartzman's procedure, as such, is his use of differentials based on calculated earnings for groups, especially those under 25 years of age, for which the data are meaningless. Some of his differentials imply that additional education actually reduces earnings.[21] The effect of defective data is accentuated by Schwartzman's use as weights to combine age-sex groups not total earnings (employment x average hours x average hourly earnings) but the product of employment and average hourly earnings. Omission of hours causes too much weight to be attached to females (for whom quality indexes rise much less than for males), the elderly, and most of all to age groups dominated by students, and magnifies the effect of bad data for these groups.

II. *Attendance per Year*

The number of school days actually attended per school year has greatly increased. The biggest reason was a huge reduction in absenteeism. In addition, the scheduled school year increased in the country as a whole, though not in big cities. In 1900, when

[17] Because education weights have least spread in the 25–34 age group, and the under-35 groups to whom they would be applied have more weight in employment than in earnings, quality indexes computed in *Why Growth Rates Differ* would tend to rise less (more) than indexes so constructed if education rose less (more) than average in the 18–35 age range. Changes in days as well as years of education must be considered to judge differential changes in education by age if one is interested in the direction of bias in my final series. My exclusion of persons under 18 from the labor force distributions greatly reduces any possible age bias. In *Sources* all persons under 25 were omitted so the bias was reversed—that is, an especially large increase in education among the young would bias my index downward. This is the opposite of Schwartzman's supposition. But it actually makes little difference whether one includes (and overweights) those 18 to 24 years of age or excludes them and thus gives them no weight at all.

[18] Schwartzman developed his own estimates of the distribution of the labor force in 1940.

[19] This is not exactly equivalent to adjustment of the earnings weights, but the difference may be unimportant in practice.

[20] Professor Schwartzman provided this explanation of his procedure, which is not fully described in his note.

[21] In the 14–19 age bracket calculated hourly earnings of males with 8 years of school exceed those of males with 9–11 or 12 years, and in the 20–24 age bracket are less for college graduates than for those with 9–11, 12, or 13–15 years. These are only the most visible oddities in the numbers.

the average schedule in a sample of northern cities was 188 days, the national average was only 144 days. The figures converged as rural and town schedules were brought up to city standards and the population became urbanized [2, pp. 30–32], and [3, pp. 88–89].

Both Schwartzman and I infer the effects of changes in days of school attended per year of school upon the quality of labor from our estimates of the effects of changes in years of education. Schwartzman obtains a smaller estimate than I (.13 in 1940–60 as against my .23 in 1950–62) because his estimate of the effects of changes in years of school is much smaller. But this is not the only difference between us.

In *Sources*, I had assumed that a 1 percent increase in the average number of days represented by a school year raised the quality of labor as much as a 1 percent increase in average years of education held by the labor force. Schwartzman cut this ratio by 30 percent, commenting:

> Some of the additional time appears to be devoted to the study of the arts and humanities, which results in the cultivation of leisure skills rather than work skills. This is the principal reason for my decision to treat additional days per school year as equivalent to 70 percent of the same number of days in additional years.

I do not find this reasoning persuasive. The causes of the increase in school days do not suggest it necessarily increased the proportion of time devoted to arts and humanities. Nor am I convinced that these subjects do less for "work skills" than others. (My procedures simply assume that differences between dates in subjects studied have not been correlated with differences in economic value of different subjects [3, p. 98].) For quite different reasons, however, in *Why Growth Rates Differ* I cut the allowance for increased days by even more than the 30 percent that Schwartzman suggests—actually by about half [3, pp. 381–83].

III. *Conclusion*

Schwartzman estimates that improvement in the educational quality of labor was 56 percent less than is shown in *Why Growth Rates Differ*. The small size of his estimate does not, as some have supposed, stem mainly from substitution of hourly for annual earnings as weights. This substitution would lower my estimate by less, I believe very much less, than 15 percent.[23] Any such change would entail a corresponding change in the definition of the educational quality index.

Three other changes in procedure that Schwartzman recommends might be considered substantive. These are use of nonfarm earnings as weights, his calculating procedure, and a 30 percent cut in the allowance for increased attendance. I have given reasons to reject all of these recommendations.

Differences between our rates that stem from time-period differences and his choice of labor force distributions cast no doubt on my estimates.

My educational quality indexes are crude; I do not wish to convey a different impression. But the points raised by Schwartzman create no such range of uncertainty as his comment suggests.

REFERENCES

1. E. F. Denison, *The Sources of Economic Growth in the United States and the Alternatives Before Us*. New York 1962.
2. ———, "Measuring the Contribution of Education (and the Residual) to Economic Growth" and "Reply," in OECD Study Group in the Economics of Education, *The Residual Factor and Economic Growth*. Paris 1964.
3. ———, assisted by J. P. Poullier, *Why Growth Rates Differ: Postwar Experience in Nine Western Countries*. Washington 1967.
4. V. R. Fuchs, *Differentials in Hourly Earnings by Region and City Size, 1959*. New York 1967.
5. D. Schwartzman, "The Contribution of Education to the Quality of Labor 1929–1963," *Amer. Econ. Rev.*, June 1968, *58*, 508–14.

[23] I argued that the percentage reduction in the Stage I estimate would be much less than 15 percent. The percentage reduction in the final estimate would be still smaller because I do not apply the "days" adjustment to persons with more than 12 years of school.

The Contribution of Education to the Quality of Labor: Reply

By David Schwartzman*

I will discuss the sources of disagreement between Edward Denison's estimate of the rate of improvement in the quality of labor between 1948 and 1965 and my own estimate for the period 1940 to 1960, which I applied to the period 1929 to 1963.

I. *Agricultural Earnings.*

The bias toward understatement in reported agricultural earnings may be offset by the other sources of error which Denison mentions. Denison's footnote at this point, however, refers to another source of downward bias in agricultural earnings as a mea-measure of labor quality: the preference of agricultural labor for farm life and resulting immobility. Denison chooses to include the influence of immobility in the 'ability' adjustment. The evidence relating to the contribution of ability to education earnings differentials is sufficiently fragmentary to permit Denison to come to this decision. My own interpretation of the available evidence is that ascribing two-fifths of the educational differentials in earnings to ability alone does not overstate ability's contribution;[1] I would not include in the two-fifths the influence of other nonschooling variables. My method of adjusting for agricultural immobility as well as for other sources of bias in agricultural earnings as reported is to use Victor Fuchs' estimates of educational earnings differentials based on a sample which excludes agricultural workers.

II. *Midpoints of Open-Ended Classes*

The choice of a uniform midpoint for the open-ended $25,000 and over class is defensible for the purpose of measuring the change in the quality of labor. Denison uses the Census's interpolated means based on Pareto curves fitted to the upper income range, which results in a larger estimated education differential. These interpolated means are acceptable for the Census's purpose, the estimation of the aggregate mean earnings; the Pareto curves correctly describe the distribution of earnings. But our purpose is to estimate education's contribution to the growth of labor quality, not aggregate mean earnings. The theories to explain the Pareto-distribution generally stress random elements, especially in relation to the upper end. If these theories are correct, random elements are chiefly responsible for differences in income in the upper range. A large part of the difference in earnings between college graduates in the upper end of the distribution and others in the labor force will be due to random elements. Denison estimates the part of the earnings of college graduates in the open-ended class which is due to schooling as the difference between the interpolated mean earnings and that of noncollege graduates within the same income class. Although the procedure may appear to be reasonable, the random element in the education differential at a high level of income is likely to be large; part of the difference between the mean earnings of college and noncollege graduates earning over $25,000 is due to random elements. Some allowance should have been made for differ-

* Professor of economics at the New School for Social Research. I am very grateful to Edward Dension for a discussion of the results which resulted in the present exchange and for comments on an earlier draft of this reply.

[1] I refer readers to Gary S. Becker [1, pp. 82–90] and Burton Weisbrod and Edward Karpoff [6, p. 495]. Weisbrod and Karpoff estimate that about one-quarter of education differentials in earnings of college graduates is due to nonschooling variables. Among all workers nonschooling variables, including ability, will be more important. Zvi Griliches, in an unpublished paper [4, pp. 34–36], suggests that estimates of income based on years of schooling are not improved by the addition of a variable representing ability because of high intercorrelation between ability and schooling. This observation does not challenge the correction for ability when what is wanted is an estimate of the improvement in the quality of labor due to education.

ences in education in the upper range of earnings. But a uniform midpoint is preferable to the Pareto-based midpoints which yield estimates that ascribe random elements in income differentials to education.

III. *Hourly versus Annual Earnings*

Denison's measure of the quality of labor is based on annual earnings; I have used Fuchs' index of average hourly earnings. The purpose is to estimate the contribution of education to the change in the quality of labor rather than to its quantity, and the use of hourly earnings is in line with this purpose. The effect of education on hours and weeks worked is an effect on quantity.

Denison has two objections to the hourly earnings index of quality. The first is that education differentials in hours and weeks worked in nonagricultural industries exceed those in agriculture. The education differentials in hourly earnings in nonagricultural industries therefore result in an understatement of education's contribution to the quality of labor in the entire economy. I agree that there may be a bias in my estimate as a result. This problem, however, is part of the general problem of the treatment of agricultural workers, which has been discussed. Uneducated agricultural workers' hourly earnings may be much lower than educated agricultural workers' hourly earnings because they are less mobile as well as less well educated.

Second, Denison objects that I assume that the output of workers who work long hours would be reduced proportionally by a reduction in hours. Productivity, of course, may have improved with the decline in hours among workers whose hours were very long. My assumption, however, is that the actual reduction in average weekly hours in the United States between 1929 and 1963 did not bring about a significant improvement in hourly productivity in the economy as a whole. Average weekly hours declined from 50 hours to 40 hours, and Irving Leveson's [5] examination of the *BLS* study on fatigue suggests that in most industries productivity does not change with hours over this range.

I turn now to the age-sex composition of the sample, the difference in time periods, and the treatment of days of schooling per year.

Age-Sex Composition of the Sample

Denison has estimated that the differences just discussed account for .17 percentage points in the difference, .32 percentage points, between his and my Stage I estimates. Denison estimates that the difference in time periods is responsible for over half the remaining difference. More precisely the time-period difference exceeds .09 percentage points. The reason the estimate cannot be stated even more precisely is the exclusion of emergency workers from the 1940 data, to which Denison refers. We are left with a difference in Stage I estimates which is less than .06 percentage points.

Denison rejects the use of earnings data for age groups under 25 years to measure the effect of education on quality within these groups. The quality of data, in his view, is inadequate. Some of his criticisms apply more to annual earnings data than to hourly earnings data, which is what I have used. Further, the criticisms apply especially to data for the group under 19 years, but Denison has excluded all persons under 25 years. In addition, the exclusion of females from the sample seems to be even less justified. In any case, the error resulting from the use of earnings of males between the ages of 25 and 64 to estimate the quality of labor among females and among males under 25 years associated with years of schooling may exceed the error from the use of data referring to these groups. Denison nevertheless says that his chief objection to my procedure is the use of differentials based on earnings for groups, 'especially those under 25 years of age, for which the data are meaningless.' He may be right: my estimate may on this account err more than his. But whatever the error is, it is responsible for less than .06 percentage points.

Time Periods

I have discussed all of the sources of disagreement between Denison's and my own Stage I estimates, except the difference in time periods. The difference, I have noted,

accounts for more than .09 percentage points of the total difference between the two estimates, .32. My data were for the years 1940 and 1960; I projected the estimated rate of change to the period 1929 to 1963, assuming no change in the rate over the whole period. Denison's estimate for 1948 to 1965 is based on data describing the educational levels in the two terminal years. Part of the difference between Denison's estimate and my own thus evidently is the result of acceleration in the rate of increase of the educational level. The difference in time periods thus is an element in the disagreement, and it is not a matter of indifference which years are used to estimate the growth rate of the quality of labor.

Days of Schooling per Year

The difference between the Stage II estimates is the result of the difference in Stage I estimates, already discussed, plus a difference in the estimates of the effect of days of schooling per year. The difference between Denison's and my estimates of the benefit from additional days per year reflects the difference in estimates of the benefit from a year of schooling; the difference in the earnings weights which produced the Stage I difference also is responsible for a large part of the difference in the estimated benefit from additional days per year. Further, the importance of the decision concerning the value to be placed on additional days depends on the estimated gain from a year of education. Because Denison's estimate of the improvement in quality resulting from an additional year exceeds my own, the choice between 100 percent (*Sources*), 70 percent (my own), or half (*Why Growth Rates Differ*) is more important for his result than for my own. If I were to adopt 50 percent instead of 70 percent, my estimate of the improvement from education, of course, would be reduced but not sufficiently to change the published estimate of .3 percent per year.

IV. *Conclusion*

Denison's Comment has been very useful in helping to find the procedural sources of difference between the estimates of the average rates of change in the quality of labor due to education. His comparison of the results of the 1948–65 measurements using different sources of data, samples, and procedures (Table 1) is especially valuable.

Before closing, I want to make three points about the disagreement concerning education's contribution to the quality of labor. In *Sources*, Denison estimated that between 1929 and 1957 the quality of labor improved at an average annual rate of .9 percent [2, p. 73]. His subsequent revision of the estimated improvement due to additional days per year reduces this estimate by .3 to .6. My own estimate for 1929 to 1963 is .3 per cent per year. A substantial difference remains in the estimates of education's contribution for a period beginning in 1929 and ending in the vicinity of 1960. Second, Denison's estimate for 1948 to 1965 exceeds my own for 1940 to 1960 by approximately .1 owing to the difference in time periods: between these two periods the rate of increase in quality due to education gained by that much. Denison's estimate for 1929–1957 compared to his estimate for 1948–1965 also indicates acceleration. Third, my assumption of constancy in the rate of improvement of the quality of labor tended to exaggerate education's contribution over the period 1929 to 1963 in view of the acceleration just noted.

REFERENCES

1. G. S. BECKER, *Human Capital*, New York 1964.
2. E. F. DENISON, *The Sources of Economic Growth in the United States and the Alternatives Before Us*. New York 1962.
3. ——— assisted by J. P. Poullier, *Why Growth Rates Differ: Postwar Experience in Nine Western Countries*. Washington 1967.
4. Z. GRILICHES, "Notes on the Role of Education in Production Functions and Growth Accounting," Conference on Research in Income and Wealth, Madison, Wisconsin, Nov. 1968 unpublished.
5. I. LEVESON, "Reduction in Hours of Work as a Source of Productivity Growth," *J. Polit. Econ.*, Apr. 1967, *75*, 199–204.
6. B. A. WEISBROD and P. KARPOFF, "Monetary Returns to College Education, Student Ability, and College Quality," *Rev. Econ. Statist.* Nov. 1968, *50*, 491–97.

Economies as an Antitrust Defense: Comment

By MICHAEL E. DEPRANO AND JEFFREY B. NUGENT*

The approach used by Oliver Williamson in the March 1968 issue of this *Review* [10], to estimate the net welfare effects of cost-reducing business mergers, provides an interesting basis upon which to challenge his conclusions and those of others who have suggested that mergers are often beneficial. Williamson calculates that "... a relatively large percentage increase in price is usually required to offset the benefits that result from a 5 to 10 percent reduction in average costs ..." and therefore concludes that "... the existence of economies of this magnitude is sufficiently important to give the antitrust authorities pause before disallowing such a merger." [10, p. 34] In this paper, Williamson's apparatus is modified and corrected, and very different conclusions are reached.

Williamson's effort follows upon a series of studies dating from the intriguing article by Arnold Harberger in this *Review* [2] and including contributions by David Schwartzman [7], [8], Harvey Leibenstein [4], C. E. Ferguson [1] and others which attempt to demonstrate theroretically and empirically that monopoly does not significantly misallocate resources or lower welfare. The empirical demonstrations have been criticized by George Stigler [9], Ruth Mack [5], and others. The recent recalculation of the welfare cost of monopoly power by David Kamerschen [3] indicates that the welfare cost of monopoly power and of mergers may have been substantially underestimated in the earlier studies. Further reflection supports these criticisms and questions the general adequacy of the approach as applied in the recent defenses of monopoly and mergers. A brief investigation of Williamson's work leads us to the view that monopoly-creating mergers are likely to impose extremely serious welfare losses, even in the face of substantial cost reductions.

* The authors are associate professor, University of Southern California and associate professor, University of Southern California and research associate, Economic Growth Center, Yale University, respectively.

The relationship between cost reductions and price is investigated in Section I and the treatment of pre-merger market power in Section II. Section III contains some comments concerning the effect of mergers on demand elasticity and income distribution as well as our conclusions.

I. *Net Welfare Effects and Price-Cost Relationships*

The Williamson approach follows earlier work in comparing the "dead-weight" loss (decreases in consumer surplus through mergers, which are not gained by the producers) to the gains to society (or, more properly, to the producers) of any cost reductions resulting from the merger.

Figure 1 is a reproduction of Williamson's diagram. In it, dead-weight loss is approximated by $\frac{1}{2}\Delta P\Delta Q$ and the cost saving is $Q_2\ (\Delta AC)$. If $Q_2\ (\Delta AC) > \frac{1}{2}\Delta P\Delta Q$, there will be a net welfare gain. Williamson rewrites this as:

$$(1)\qquad \frac{\Delta AC}{AC} - \frac{k}{2}\eta\left(\frac{\Delta P}{P}\right)^2 > 0$$

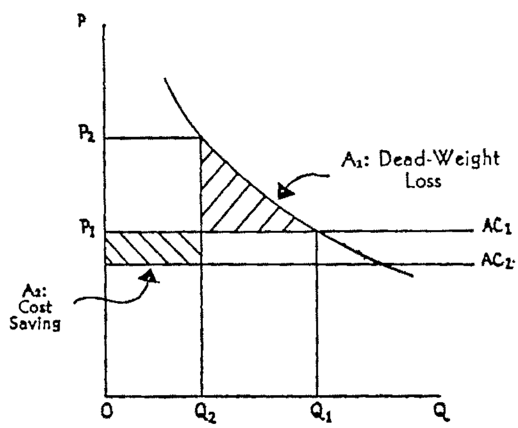


FIGURE 1

where η is the absolute value of the elasticity of demand and k is "an index of pre-merger market power" and equals P_1/AC_1. His discussion gives the impression that AC and P in equation (1) have the subscript 1—the pre-merger price and average cost—and that his measure of elasticity is $(\Delta Q/\Delta P)$ (P_1/Q_2). Inequality (1) indicates "... that if the decimal fraction reduction in average costs exceeds the square of the decimal fraction increase in price multiplied by one-half k times the elasticity of demand, the allocative effect of the merger is positive" [10, p. 22]. So, for example, with a 10 percent increase in price after the merger, a cost reduction of more than one percent will lead to increases in welfare, if the elasticity of demand is 2. A cost reduction of only one-half of one percent would be sufficient to neutralize the impact of such a price increase if the elasticity of demand is 1.

Williamson's calculations of the price and cost changes which would leave welfare unchanged are correct if $k=1$, i.e., $P_1=AC_1$.[1] What Williamson fails to realize is that the cost reductions which he considers would probably be associated with much larger price increases than those which he indicated can easily be offset! Costs and prices are obviously not independent of each other. In the case of mergers leading to monopoly, for example, cost reductions on the order of 1, 5, or 10 percent would be associated with price increases *many times* greater than those whose negative effects would be overcome by the cost reductions.

The horizontal average cost curve used by Williamson (and also Harberger [2], Schwartzman [7], [8], Leibenstein [4], and others) would indicate constant long-run costs, and the interesting comparisons are, of course, between long-run equilibrium positions. In this case, then, $MC=AC$. For a monopoly which is not constrained from seeking its best profits, $MC=MR$. However, as Joan Robinson [6] has shown,

$$P = \frac{\eta}{\eta - 1}(MR).^2$$

Therefore,

$$P_2 = \frac{\eta}{\eta - 1}(AC_2).$$

If

$$P_1 = AC_1$$

then

$$\frac{P_2}{P_1} = \frac{\eta}{\eta - 1}\frac{AC_2}{AC_1}.$$

Converting this expression into one in terms of percentage changes in order to compare it with expression (1) (which is Williamson's measure of the welfare tradeoff relationships between price and cost changes), we obtain

$$\frac{\Delta P}{P_1} = \frac{1 - \eta\left(\frac{\Delta AC}{AC_1}\right)}{\eta - 1} \tag{2}$$

Assume for a moment η equals 2 (which Williamson suggests "... is probably a reasonable upper bound"). It is correct that a one percent decrease in costs is sufficient to offset the welfare effects of a 10 percent increase in prices, as equation (1) shows. However, equation (2) indicates that a monopoly-creating merger which brought a one percent cost reduction would lead to a price increase of 98 percent. Likewise, a 9 percent cost reduction can offset welfare effects of a

[1] The case of $P_1>AC_1$ is discussed in Section II. Williamson has made a serious error in his discussion of the case in which some monopoly power exists prior to merger.

[2] For a demand curve which does not have a constant elasticity, this elasticity measure and Williamson's are not quite the same. His original measure from (1) above is $(\Delta Q/\Delta P)$ (P_1/Q_2), an approximation to the elasticity between the old and new points on the demand curve and his recent modification of that measure, given by his equation (3′) in [11], is the elasticity at the old point. Although in our present discussion we shall refer only to Williamson's original formulation, none of our conclusions would be affected if we used his modified formula instead. His original measure and ours can be made comparable in several ways. One simple way would be to replace $\eta(\Delta P/P)^2$ in (1) with $\eta_2(\Delta P/P_1)$ $(\Delta P/P_2)$, where η_2 is the elasticity at the new point. This would show that the different elasticity measures do not seriously affect our conclusions. In these equations η is the absolute value of the elasticity.

TABLE 1—RELATIONSHIP BETWEEN COST REDUCTIONS AND PRICE INCREASES[a]

Percentage Cost Reduction	Elasticity of Demand					
	2		1.5		1.1	
	Percentage Price Increase					
A	B	C	B	C	B	C
.12	3.4	99.76	3.92	199.6	6.18	998.75
.25	5.0	99.50	5.76	199.2	9.09	997.25
.50	7.0	99.00	8.07	198.5	12.72	994.50
1.00	10.0	98.00	11.53	197.0	18.18	989.00
2.00	14.14	96.00	16.30	194.0	25.21	978.00
3.00	17.32	94.00	19.97	191.0	31.49	967.00
4.00	20.00	92.00	23.00	188.0	36.36	956.00
5.00	22.36	90.00	25.78	185.0	40.65	945.00
10.00	31.62	80.00	36.48	170.0	57.48	890.00
20.00	44.27	60.00	51.04	140.0	80.48	780.00
30.00	54.76	40.00	63.14	110.0	99.55	670.00
40.00	63.24	20.00	72.92	80.0	114.97	560.00
50.00	70.71	0.00	81.53	50.0	128.55	450.00
60.00	77.45	−20.00	89.30	20.0	140.80	340.00
70.00	83.66	−40.00	96.46	−10.0	152.09	230.00
80.00	89.44	−60.00	103.12	−40.0	162.60	120.00

[a] Column (A): Cost reductions
Column (B): Price increases whose welfare effects could be offset by the cost reductions
Column (C): Price increases that would result from monopolization with the given cost reductions

30 percent increase in prices, but if costs fall by 9 percent, prices would increase by 82 percent. The related cost decreases and price increases forthcoming from such mergers, therefore, would normally give very strong *negative* welfare effects.

At smaller levels of elasticity, it is correct that even smaller cost reductions are required to offset the welfare effects of a given price increase. However, lower elasticities would lead to even higher price increases for any given cost reduction resulting from a merger.[3] Table 1, Column A, lists various cost reductions assumed to result from merger and Columns C list the price increases which would be brought about by the same monopoly-creating merger, for various elasticities. The B Columns give the percentage changes in prices which could be offset by the corresponding cost reductions in Column A for various elasticities. For example, an elasticity of 1.1, according to Williamson's equation (1), would indicate that a cost reduction of .50 percent would be sufficient to eliminate the negative effects of a 12.72 percent price increase. Equation (2) shows that a .50 percent cost reduction at that elasticity would be associated with approximately a 10-fold increase in price after the merger. A comparison of Columns B and C shows that only very large cost reductions would be sufficient to offset the price rise that would accompany monopolization of the industry. When the demand elasticity is 2, the cost reduction must be over 20 percent; with an elasticity of 1.5, the cost reduction would have to be over 40 percent; and at an elasticity of 1.1, a cost reduction greater than 70 percent is necessary!

Of course, mergers need not lead to monopoly. For mergers which gain less than monopoly power, the price rise would be lower than that indicated by equation (2), but it is also unlikely in our economy that the demand elasticity would be as high as 2.

[3] Of course, a monopoly firm would operate in an area where the absolute value of η is greater than 1.

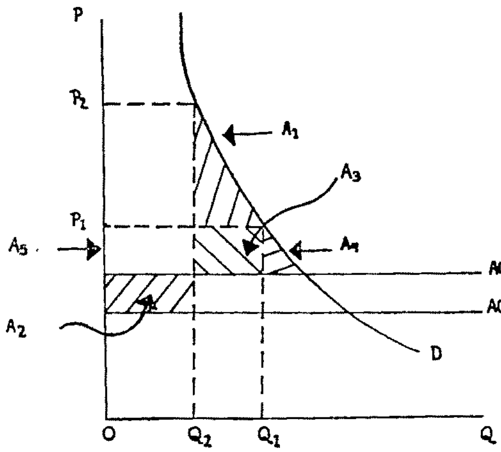


FIGURE 2

With lower elasticities, even small degrees of post-merger market power would be likely to bring price increases considerably higher than could be overcome by cost reductions. Equation (1) shows that the cost reductions, which are needed to offset the welfare effects of a given price increase, fall in a constant proportion to decreases in elasticity. Equation (2) indicates that price increases associated with a given cost reduction rise *geometrically* as elasticities are reduced. With low elasticities even a very minor merger would need unrealistically large cost reductions to avoid negative welfare effects!

II. *Pre-Merger Market Power*

If some degree of market power exists before the merger, then a given post-merger firm size would have less effect on price than if pure competition prevailed previously. However, the relation between the price rise and cost reduction necessary to bring welfare gains is seriously understated by the Williamson formulation, equation (1). Williamson recognizes the possible existence of pre-merger market power by introducing "k" as an index of such power. When an industry is purely competitive, in long-run equilibrium, $P=AC$, and $k=1$, but Williamson uses k only for mathematical purposes to reduce his equation to the simple form given by (1). He overlooks the fact that the existence of pre-merger market power will call for a revision of his original formula: $Q_2(\Delta AC) > \frac{1}{2}\Delta P\Delta Q$.

Consider Figure 2. The total returns to society of the pre-merger output (Q_1) is the total area above AC_1, i.e., the consumer surplus plus areas A_3 and A_5. To confirm this, compare the relationships of some pre-merger market power to monopoly in Figure 2, and of pure competition to monopoly in Figure 1. In Figure 2, the areas A_1, A_3, and A_4, taken together represent the dead-weight loss that was represented by A_1 in Figure 1. If pure competition prevails before merger, the monopoly-creating merger would involve a welfare loss of A_1, A_3, and A_4, but a gain of A_2. If pre-merger market power had existed before the merger, then this would have eliminated A_4, and *only* A_4, from total welfare. Therefore, A_1 and A_3 are both eliminated by a merger when market power exists before the merger. Williamson confuses A_1 in Figure 1 with A_1 in Figure 2, and so treats only A_1 in Figure 2 as lost. Correctly formulated, the condition for a net welfare gain should be $(\Delta AC)Q_2 > \frac{1}{2}\Delta P\Delta Q + (P_1 - AC_1)\Delta Q$, which can be rewritten:

$$(3) \qquad \frac{\Delta AC}{AC_1} > \frac{k}{2}\eta\left(\frac{\Delta P}{P}\right)^2 + (k-1)\eta\frac{\Delta P}{P}$$

Because inequality (3) includes a (positive) term that was not contained in (1), it might appear that the merger would bring about a larger loss in welfare if market power, instead of perfect competition, existed before the merger. That this is incorrect can be seen by noting that the term $\frac{1}{2}\Delta P\Delta Q$ which, in the case of pure competition, would include A_1, A_3 and A_4, includes only A_1 in the case of pre-merger market power. If competition prevailed before the merger, the term $(P_1 - AC_1)\Delta Q$ would not be needed, but $\frac{1}{2}\Delta P\Delta Q$ would measure a much larger area: the sum of A_1, A_3, and A_4.

Williamson's neglect of A_3 in his formulation leads to a serious understatement of the cost reductions needed to offset the welfare effects of price increases in the presence of pre-merger market power. The extent to which Williamson understates the cost reductions necessary to offset the price in-

TABLE 2—PERCENTAGE COST REDUCTIONS [$(\Delta AC/AC)\times 100$] SUFFICIENT TO OFFSET PERCENTAGE PRICE INCREASES $(\Delta P/P)\times 100$ FOR SELECTED VALUES OF η AND k USING WILLIAMSON FORMULA (1).

η	2		1		½	
$(\Delta P/P)\times 100$	$k=1.5$	$k=1.1$	$k=1.5$	$k=1.1$	$k=1.5$	$k=1.1$
5	.375	.275	.188	.138	.084	0.690
10	1.500	1.100	.750	.559	.375	.275
20	6.000	4.400	3.000	2.200	1.500	1.100
30	13.500	7.900	6.750	4.950	3.375	2.475

creases can be determined by comparing Tables 2 and 3. Both tables present required cost reductions for $k=1.1$ and $k=1.5$. However, Table 2 is based on Williamson's formulation, equation (1) above, while Table 3 is based on the correct formulation, expression (3). Note that the results using Williamson's formulation are very similar to those of his Table 1 [10, p. 23], whereas those of Table 3 are significantly higher. For example, when $k=1.5$ and $\eta=2.0$, Williamson's formula indicates that the welfare effects of a 10 percent price increase can be offset with a 1.5 percent cost reduction. Equation (3) indicates that an 11.5 percent cost reduction would be needed! Since antitrust authorities usually direct their attention to mergers of firms which already have significant market shares, Williamson's gross understatement of the cost reductions necessary to offset price increases when $k>1$ is an important error.

If equation (2) in Section I were revised to incorporate $P_1>AC_1$, $(k>1)$, it would show, as one might expect, that with the existence of pre-merger market power, a given cost reduction leads to smaller price rises than when there is pure competition before the merger. On the other hand, equation (3) and Table 3 show that the larger k is, the smaller is the price rise whose effects can be overcome with a given cost reduction. Tables similar to Table 1 could be constructed to indicate the specific cost reduction necessary for a monopoly-creating merger to bring net welfare benefits in cases in which k is greater than unity. Such calculations would, of course, show that the greater the value of k, the smaller the cost reductions required to bring net welfare gains. Nevertheless, for the levels of k and η which are normally found in our economy, very large cost reductions would be necessary. Certainly, the 5–10 percent indicated by Wil-

TABLE 3—PERCENTAGE COST REDUCTIONS [$\Delta(AC/AC)\times 100$] SUFFICIENT TO OFFSET PERCENTAGE PRICE INCREASES $(\Delta P/P)\times 100$ FOR SELECTED VALUES OF η AND k USING CORRECT FORMULA (3).

η	2		1		½	
$(\Delta P/P)\times 100$	$k=1.5$	$k=1.1$	$k=1.5$	$k=1.1$	$k=1.5$	$k=1.1$
5	5.375	1.275	2.688	.637	1.344	.319
10	11.500	3.100	5.750	1.550	2.875	.775
20	26.000	8.400	13.000	4.200	6.500	2.100
30	43.500	15.900	21.750	7.950	10.875	3.970

liamson as calling for "pause" on the part of antitrust authorities is far from what is indicated here. The kinds of cost reductions our calculations suggest as needed are of a size that mergers are extremely unlikely to bring. If such economies exist, the industry involved would likely be a "natural monopoly" and already regulated as such.

III. *Other Considerations and Conclusion*

The previous sections followed Williamson in using an elasticity of given value. However, most demand curves are likely to have different elasticities at different prices—generally higher elasticities at higher prices. Therefore, if a merger does not shift the demand curve, (through advertising, etc.) but simply moves along the curve to a higher price, the elasticity of demand with respect to price would normally rise. Revising Williamson's formula slightly to include the impact of the merger in raising the demand elasticity, one would find that greater cost reductions are needed to offset the effects of given price rises. However, for reasonable values of k, and η, these differences are not as striking as those presented in Sections I and II, and therefore will not be presented here.

Finally, the treatment by Williamson (and others who have defended mergers) of the income redistribution effects of mergers calls for a brief comment. Indeed, Williamson recognized their importance by suggesting that his formulation, (1) could be weighted to take into consideration any income distribution effects, and concedes ". . . attaching even a slight weight to income distribution effects can sometimes influence the overall valuation significantly." [10, p. 28] However, he prefers that antitrust authorities concern themselves with the "allocation" problem instead of dealing with income distribution objectives. He argues that since ". . . antitrust is an activity better suited to promote allocative efficiency than income distribution objectives (the latter falling more clearly within the province of taxation, expenditure, and transfer payment activities), such income distribution adjustments might routinely be suppressed." [10, p. 28]. Williamson (and others who take this position) overlook the fact that income distribution is, itself, a very crucial part of the allocative process. The tax or the transfer activities which would seek to mitigate or compensate for the effects of mergers on income distribution would *also* lessen income differentials which bring about desirable allocations of resources! Taking monopoly gains from producers and giving them back to consumers is quite a different story than cutting the incomes of producers who are making their gains by producing items of extreme scarcity. The high incomes of the latter are *inducements* to greater production of things society values highly. The monopoly gains are from contrived scarcities. This is not to suggest that the goal of equity in income distribution in and of itself is necessarily poor economics. Still, it must be recognized that when one wishes to redistribute high incomes which are themselves the result of production of those goods which society values highly, a definite choice between the goals of equity and efficiency must be made. The elimination of income differences resulting from market power brings no such unhappy choice. Lessening market power both lessens income inequities related to the market power *and*, under most of the conditions likely to exist in our economy, increases allocative efficiency. The use of tax, expenditure, and transfer activities to eliminate income differentials cannot easily be made to discriminate between the income differentials created by market power and those contributing toward greater allocative efficiency. If those caused by market power could be eliminated, of course, a society might still rationally decide to sacrifice some allocative efficiency to gain more equitable income distributions.

Williamson's paper, therefore, has provided us with an excellent vehicle to draw conclusions widely at variance with his. Rather than suggest "pause" on the part of antitrust authorities, even in the face of prospective cost reductions, our comments suggest that the required cost reductions would have to be so large that they are very

unlikely to be forthcoming. Rather than concerning themselves with what would seem to be a rare occasion in which a cost-reducing merger might increase producers' welfare more than consumers' loss, it would seem that more rigorous activity on the part of antitrust authorities would be likely to serve both the goal of increasing net economic welfare, and that of equity in income distribution. However, if the antitrust authorities insist on pursuing investigations similar to those proposed by Williamson, it would seem imperative, at the very least, that they use the correct formula!

References

1. C. E. Ferguson, *A Macroeconomic Theory of Workable Competition*, Durham, N. C., 1964.
2. A. C. Harberger, "Monopoly and Resource Allocation," *Amer. Econ. Rev., Proc.*, May 1954, *44*, 77–87.
3. D. R. Kamerschen, "An Estimation of the Welfare Losses from Monopoly in the American Economy," *Western Econ. J.*, summer 1966, *4*, 221–36.
4. H. Leibenstein, "Allocative Efficiency versus X-Efficiency," *Amer. Econ. Rev.*, June 1966, *56*, 392–415.
5. R. P. Mack, "Discussion," *Amer. Econ. Rev., Proc.*, May 1954, *44*, 88–89.
6. J. Robinson, *The Economics of Imperfect Competition*, London 1933.
7. D. Schwartzman, "The Effect of Monopoly on Price," *J. Polit. Econ.*, Aug. 1959, *67*, 352–62.
8. ———, "The Burden of Monopoly," *J. Polit. Econ.*, Dec. 1960, *68*, 627–30.
9. G. Stigler, "The Statistics of Monopoly and Merger," *J. Polit. Econ.*, Feb. 1956, *64*, 33–40.
10. O. E. Williamson, "Economies as an Antitrust Defense: The Welfare Tradeoffs," *Amer. Econ. Rev.*, Mar. 1968, *58*, 18–36.
11. ———, "Economies as an Antitrust Defense: Correction and Reply," *Amer. Econ. Rev.*, Dec. 1968, *58*, 1372–76.

Economies as an Antitrust Defense: Reply

By OLIVER E. WILLIAMSON*

The Comment by Michael DePrano and Jeffrey Nugent is in three parts. I conclude with respect to the first part that the pricing formula which they employ is operationally irrelevant and that this section (text, tables, formula) can be disregarded altogether. Section II of their Comment identifies an omission in my original treatment of the tradeoff issue that properly deserves to be corrected. Not content to make the correction, however, they go on vastly to exaggerate its significance by their assignment of parameter values. After providing an alternative derivation of the correction, I show that the original argument is not upset if parameter values in the relevant range are selected.

The third part of their Comment is mainly concerned with the income distribution issue. I had originally planned to respond to this part also, but the obscurity of the discussion has dissuaded me. (Consider the following: "Taking monopoly gains from producers and giving them back to producers is quite a different story than cutting the incomes of producers who are making their gains by producing items of extreme scarcity." [4, p. 952]) I merely point out that, although concern over distributional effects is entirely legitimate, to rely on the antitrust agencies and the courts to implement such a program encounters formidable problems; and to express a preference for taxes, transfers, and expenditure programs to remedy distributional questions obviously does not imply that any of these activities is free from side effects (it only reflects a judgment of net consequences) I have discussed both of these matters elsewhere [16].[1]

I. *The Price-Making Process*

Recall that I postulated a merger that simultaneously had cost reducing and market power consequences [14]. DePrano and Nugent accept the cost reduction condition and take as their standard for the creation of market power the transformation of a previously competitive industry into a thoroughly monopolistic one.[2] This is somewhat extreme and scarcely implied by my argument, but let us play the game their way.

They then inform us that monopolies set price according to the formula that $P=MC/[1\text{-}(1/\eta)]$, and go on to display their equation (2) which shows a "causal" association between price increases and cost reductions. While I concede that a not uncommon introduction to the concept of monopoly pricing often begins with the conventional formula noted above, most such discussions, unlike DePrano and Nugent's, go on to consider the limitations of this formula—which most students of industrial organization would regard as substantial. The reason, of course, is that for the formula to have significance requires that the monopoly be free from the threat of entry. Otherwise the monopolist will price with a view to the effect that his margins will have on entry.

The argument is hardly novel. Numbered among its early expositors are Alfred Marshall [6, pp. 269–80] and J. B. Clark [2, pp. 25–30]. More recent contributors include Tibor Scitovsky [9, pp. 22–23], Joe Bain [1], Sylos Labini [10], and Franco Modigliani [7]. Scitovsky expresses it as follows: "The individual price maker has to meet two forms of competition: the actual competition of his established rivals and the threat of competition from newcomers to his market. . . . Of the two . . . the threat from newcomers and restraints on their entry to the market

* University of Pennsylvania.

[1] For an interesting study in contrast to the DePrano-Nugent view on cost saving monopolies, see [5, pp. 562–68].

[2] Their qualification that "Of course, mergers need not lead to monopoly" [4, p. 949] is not very useful since they give no specific indication of how significantly their previous results would be altered. They leave the impression, however, that the basic argument is not, in any essential respect, greatly affected.

are by far the more important from the price maker's point of view." [9, pp. 22–23]. De Prano and Nugent, nevertheless choose to disregard the entry question.

As indicated above, I regard the standard that a cost saving merger transforms an industry from competitive to monopolistic structure as implausible. The unexpressed but crucial assumption that entry is simultaneously barred does not even meet this criterion. Combining such assumptions predictably yields absurd results—as is evidenced by such statements as "equation (2) indicates that a monopoly-creating merger which brought a 1 percent cost reduction would lead to a price increase of 98 percent!" [4, p. 949].

Consider instead the following description of the price-making process. Firms A and B actively compete for a common group of customers but (say by reason of spatial separation) are partially insulated from other firms in the industry. Assume that collusion is illegal (or in any case is difficult to sustain) and that, lacking a "justification" for merger, the two firms are prohibited from combining. Although recognized interdependence will tend to attenuate aggressive rivalry, residual competition between them can nevertheless be expected. Prices will therefore remain below joint profit maximizing levels. Now, however, suppose that a significant source of cost savings develops which can only be realized by their joint operations. If the combination is permitted, such competition as had previously existed between them will naturally vanish. Unless, however, entry into their market is "effectively blockaded" (in the sense expressed by Bain [1, p. 22]; that is, other firms will neither transport into nor otherwise invade their market), the monopoly price will not obtain.

The question can thus be put: What, realistically, is the likelihood that an effectively blockaded entry condition will result from the combination? John Wenders concludes, I think correctly, that effectively blockaded entry falls in the null set [11]. Short of this extreme, what alternative effects might be anticipated? Judging from Bain [1, p. 170] and other evidence on price to cost margins [3], an increase of well less than 10 percent, which reflects partial but incomplete insularity from existing and potential rivals, is probably to be expected.

The pricing formula corresponding to this description of behavior can be obtained by expressing the firm's maximization problem as

$$\text{(1)} \qquad \begin{aligned} &\text{maximize:} \quad \pi = PQ - C(Q) \\ &\text{subject to:} \ P \leqq \overline{P}, \end{aligned}$$

where P is price ($P=(Q)$, $P'<0$), Q is output, $C(Q)$ is cost, and $\overline{P}$ is the price at which entry can reasonably be anticipated.[3] The derived pricing formula is:

$$\text{(2)} \qquad P = MC\Big/\left[1 - \frac{1}{\eta}\left(1 - \frac{\lambda}{Q}\right)\right],$$

where λ is the Lagrange multiplier and reflects the shadow price of the constraint. Only if the constraint is redundant, in which case $\lambda = 0$, will the conventional monopoly pricing formula obtain. Also note that the statement that a monopolist will always operate in the region where demand is elastic, which DePrano and Nugent dutifully repeat [4], is not implied when the constraint is binding.[4] In circumstances where prices are "margin restricted," demand elasticity is determined rather than determining.

Other market scenarios could also be described.[5] Among those that qualify as reason-

[3] For a probabilistic formulation, see [12, pp. 124–27]. A more recent and extensive development of a probabilistic approach to the entry question is given in [5].

[4] The elasticity referred to in equation (2) is that of the demand curve $P=P(Q)$ evaluated at $\overline{P}$, not that of the truncated demand relation at this point.

Note that the maximal margin ($\overline{P}$-AC) is strictly independent of demand elasticity only where the price that would attract established outsiders to ship into the region is lower than the price that would attract new firm formation within the territory, and only if the marketing expense (including transportation) of outsiders is subject to negligible scale economies with respect to the marginal sales. Otherwise, the optimal value of $\overline{P}$ is not independent of demand elasticity: *ceteris paribus*, the higher is η, the lower is $\overline{P}$. But the value of $\overline{P}$ thus derived will not, in general, be the same as the optimum price given by DePrano and Nugent in their conventional monopoly pricing formula.

[5] The multi-market situation in which, as a result of the merger, market occupancy reaches market power

able, the DePrano-Nugent treatment does not emerge as an obvious candidate. Indeed, although for many of these scenarios the post-merger price to cost margin will exceed the pre-merger spread, I would expect, in general, that mergers resulting in economies would produce *price decreases*—which does great violence to their curious notion of causality. I postulated the price increase in the naive model simply to expose the tradeoff issue. DePrano and Nugent carry it to fanciful extremes.

There is nothing novel in the proposition that market power does not imply monopoly or that potential competition does not require prior market occupancy. DePrano and Nugent override both, which permits them to reach "very different conclusions." They will not be surprised, I trust, if their invitation to adopt these is not widely accepted.

II. *Pre-Existing Market Power*

The welfare function on which my original formulation of the "naive" model was (intendedly) based is the same as that which is conventionally used for partial equilibrium welfare valuations (see, for example [13]); to wit, welfare is expressed as $W=(TR+S)-(TC-R)$, where, under appropriate restrictions, the terms in the first set of parenthesis reflect social benefits (total revenue plus consumers' surplus) and those in the second reflect social costs (total pecuniary costs less intramarginal rents). Assuming that R is negligible, this can be restated as $W=(TR-TC)+S$, or inasmuch, as $TR-TC$ reflects profits, as $W=\pi+S$, where π denotes profits. Any change in welfare is then given by

$$\Delta W = (\pi_2 - \pi_1) + (S_2 - S_1).$$

This result is perfectly general. The problem comes in applying it to particular circumstances. My original translation in the naive model is adequate with respect to π_2, S_2, and S_1, but it incompletely characterizes π_1. Thus, whereas most of π_1 is accounted for in the original tradeoff expression (equation (1) in [14, p. 22]), a deadweight loss element has been omitted.[6] Hence, as De Prano and Nugent point out, this expression is strictly correct only if pre-existing market power is negligible ($\pi_1=0$).

Possibly the naive model should be scrapped on this account, but I would suggest that it be regarded as representing the case where pre-existing market power is negligible. The case where nonnegligible premerger power exists can then be treated as one of several qualifications to this naive formulation. Three issues are thereby presented: 1) What is the correct tradeoff expression when pre-existing market power obtains? 2) How significantly are the results of the naive model changed when plausible values of the market power parameter are assigned? 3) How important is this correction in relation to other qualifications to the argument already covered in the original discussion [14, pp. 24–32], [15]?

Again let k be the pre-existing market power parameter, where $k=P_1/AC_1$. Employing the framework suggested above, it can be shown that ΔW can be expressed as

$$(3) \quad \Delta(AC)Q_2 - [\tfrac{1}{2}\Delta P + (k-1)AC_1]\Delta Q.$$

The parallel test expression to equation (3′) [15, p. 1373] is whether

$$(4) \quad \frac{\Delta(AC)}{AC_1} - \left[\frac{1}{2}k\left(\frac{\Delta P}{P_1}\right) + (k-1)\right]\eta\frac{\Delta P}{P_1}\cdot\frac{Q_1}{Q_2}$$

proportions in a few but not in all of the markets is one such example. Differential price to cost margins between markets might therefore be expected. If economies are realized on all transactions but margin increases occur for only a few, the welfare valuation must be weighted accordingly.

[6] π_1 is given by $P_1Q_1-(AC_1)Q_1$ or, letting $P_1=kAC_1$, by $(k-1)Q_1(AC_1)$. This can be expressed as the sum of $(k-1)Q_2(AC_1)$ and $(k-1)\Delta Q(AC_1)$. The first of these terms is accounted for by my equation (1), but the second is omitted. The ratio of omitted to included is thus given by $\Delta Q/Q_2$. For demand elasticity in the neighborhood of unity and for price increases in range of 0 to 10 percent (where, for reasons given below, 10 percent may be regarded as a rough upper bound), the value of $\Delta Q/Q_2$ will be approximately equal to the percentage price increase. Thus, as indicated in the text, most but not all of π_1 is accounted for by equation (1), [14, p. 22].

TABLE 1—PERCENTAGE COST REDUCTIONS $[(\Delta AC/AC_1)\times 100]$ SUFFICIENT TO OFFSET PERCENTAGE PRICE INCREASES $[(\Delta P/P_1)\times 100]$ FOR SELECTED VALUES OF η AND $k=1.00$ AND 1.05

$\Delta P/P_1 \times 100$	$\eta=2$		$\eta=1$		$\eta=\frac{1}{2}$	
	$k=1.00$	$k=1.05$	$k=1.00$	$k=1.05$	$k=1.00$	$k=1.05$
5	.26	.78	.12	.38	.06	.19
10	1.05	2.15	.50	1.03	.24	.50
20	4.40	6.82	2.00	3.10	.95	1.48
30	10.35	14.28	4.50	6.21	2.10	2.90

is greater or less than zero.[7]

The $(k-1)$ term in this expression was omitted from my original test relation. It reflects the neglected part of π_1 mentioned above. The degree to which pre-merger market power requires a qualification in the results obtained from the naive model obviously varies directly with the value of k. DePrano and Nugent treat the range of k from 1.1 to 1.5, and presumably regard it as relevant. Unfortunately, however, they provide neither a theoretical nor empirical basis for postulating such a range. Although their intuition in this instance may be good, one learns to be wary.

Consider Bain's treatment of entry barriers in the "very high," "substantial," and "moderate or low" categories [1, p. 170]:

> It is hazardous to assign any absolute values to the entry barriers corresponding to these three rankings, but the very roughest guess would be as follows: (1) that in the "very high" category, established firms might be able to elevate price 10 percent or more above *minimal* costs while forestalling entry; (2) that with "substantial" barriers, the corresponding percentage might range a bit above or below 7 percent; (3) that in the "moderate to low" category the same percentage will probably not exceed 4, and will range down to around 1 percent. (Emphasis added.)

Inasmuch as we are postulating that the merger will result in both cost reductions and an increase in price above the pre-merger level, and if a 10 percent post-merger price to cost margin should be regarded as considerable, it is difficult to imagine that values of k in the range 1.1 to 1.5 are often to be expected. Thus let $P_2=\theta P_1$, where $\theta>1$, and $AC_2=\gamma AC_1$, where $0<\gamma<1$. Now if $P_2/AC_2=m$ it follows that $k=P_1/AC_1=(\gamma/\theta)m$. Maximal values for k will obtain when γ and m are assigned as large values as can plausibly be imputed to them and when θ is made as small as possible. If a value of m as great as 1.10 is regarded as a rough upper bound (this would place the merged firm in the high entry barrier category), and if θ and γ are both pressed near to unity (say $\theta=1.02$ and $\gamma=.98$, which might be regarded as threshold values for each), the resulting maximal value for k is (roughly) 1.06. Evidently the range 1.1 to 1.5 which DePrano and Nugent explore is, for the most part, operationally irrelevant.

Instead I would expect that values of k in the neighborhood of 1.00 would be typical, that values as high as 1.05 would obtain occasionally, but that values of k as great as 1.10 would be rare. I show in Table 1 the percentage cost reductions sufficient to offset various percentage price increases for selected demand elasticities (assumed to be constant in the relevant range) and values

[7] This expression differs slightly from DePrano and Nugent's expression (3) [4, p. 950] because they decline to employ the correction that I suggest in [15, pp. 1272–73]. This affects their numerical results slightly, but the major cause for discontent with this part of their argument is their assignment of implausible values to k.

of k of 1.00 and 1.05.[8] Inspection of the table immediately reveals that the differences which obtain when values of k as large as 1.05 are introduced (or even for $k=1.1$; see the DePrano-Nugent Table 3) are, while not unimportant, hardly sufficient to alter previous conclusions drastically. Thus the introduction of pre-merger market power in plausible amounts does not, by itself, upset the the following proposition derived from the naive model: "a merger which yields nontrivial real economies must produce substantial market power and result in relatively large price increases for net allocative effects to be negative" [14, p. 23]. If De Prano and Nugent are really serious in their statement that they are interested in "levels of k and η which are normally found in our economy" presumably they would agree.

This is not to suggest that the results of the naive model cannot be upset under any reasonable set of circumstances. I expect, however, that the more important qualifications involve enforcement expense, timing, incipiency, weighting, income distribution, extra-economic political objectives, technological progress, and the effects of monopoly power on managerial discretion—all of which are discussed elsewhere [14, pp. 24–32], [15], [16].[9] Pre-existing market power may be associated with some of these, but that is evident from the original argument.[10] The DePrano-Nugent Comment is concerned with the local effects of pre-merger market power by itself. As shown above, this does not have the devastating consequences that they would impute to it.

REFERENCES

1. J. BAIN, *Barriers to New Competition*, Cambridge, Mass. 1956.
2. J. B. CLARK, *The Control of Trusts*, New York, 1914.
3. N. R. COLLINS AND L. E. PRESTON, *Concentration and Price-Cost Margins in Manufacturing Industries*, Berkeley 1968.
4. M. E. DEPRANO AND J. B. NUGENT, "Economics as an Antitrust Defense; Comment," *Amer. Econ. Rev.*, *59*, Dec. 1969, 947–53.
5. M. I. KAMIEN and N. L. SCHWARTZ, "Limit Pricing and Uncertain Entry," W. P. P-5-69-7, Graduate School of Industrial Administration, Carnegie-Mellon Univ., Aug. 1969.
6. A. MARSHALL, "Some Aspects of Competition," in A. C. Pigon, ed., *Memorials of Alfred Marshall*, New York 1956, pp. 256–94.
7. F. MODIGLIANI, "New Developments on the Oligopoly Front," *J. Polit. Econ.* June 1958, *66*, 215–32.
8. R. A. POSNER, "Natural Monopoly and

[8] If, for reasons given in the text, a value of m as great as 1.10 is to be regarded as a rough upper bound, then only the first two rows in the table are relevant when $k=1.00$, and only the first row when $k=1.05$. I show values of $\Delta P/P_1$ as great as 30 only for purposes of illustrating how great a price increase is required to offset a cost reduction as large as 5 percent (if the assumptions of the argument were to hold), not because I regard price increases in this degree as realistic.

[9] Also, I should note, the second-best qualification in [15, p. 1373] should stipulate that the significant relations with other sectors in which distortions exist are mainly ones of substitutability rather than complementarity—which ordinarily seems reasonable. A still further qualification concerns the product differentiation issue. Any predictable change in product variety attributable to the merger should appropriately be reflected in the welfare valuation. Inasmuch, however, as product variety changes could go either way (an increase or decrease in product variety is possible), since neither is apt to be evident *ex ante*, and as such effects as do obtain are likely to be small, it seems reasonable to proceed, for the present, at least, as though product variety will remain unchanged. This is the usual, if implicit, assumption in most welfare loss valuations, including [14]. For a recent contribution to the product differentiation literature that might, were the matter to be regarded as operationally significant, eventually be brought to bear on these issues, see [17].

[10] For example, weighting. Here the argument is that while economies resulting from the merger are apt to be limited to the combining firms, the market power effects may spread more generally. If the cost saving is realized on production Q_2 while the price increase is effective across sales Q_T, the dead-weight loss shown in the naive model should be multiplied, roughly, by the ratio Q_T/Q_2. (This assumes that the firm demand curve can be expressed in a pro-rata market share relation to the total.) Multiples of five or more are not wholly unrealistic. Such adjustments tend to swamp the De Prano-Nugent correction, although this is not to suggest that the latter should be neglected. For a weighting correction with countervailing effects, see fn. 5.

Its Regulation," *Stanford Law Rev.*, Feb. 1969, *21*, 548–643.
9. T. SCITOVSKY, *Welfare and Competition.* Chicago 1951.
10. P. S. LABINI, *Oligopoly and Technical Progress*, Cambridge, Mass. 1962.
11. J. T. WENDERS, "Entry and Monopoly Pricing," *J. Polit. Econ.*, Oct. 1967, *75*, 755–60.
12. O. E. WILLIAMSON, "Selling Expense as a Barrier to Entry," *Quart. J. Econ.*, Feb. 1963, *77*, 112–28.
13. ———, "Peak-Load Pricing and Optimal Capacity Under Indivisibility Constraints," *Amer. Econ., Rev.* Sept. 1968, *56*, 810–27.
14. ———, "Economies as an Antitrust Defense: The Welfare Tradeoffs," *Amer. Econ. Rev.*, Mar. 1968, *58*, 18–36.
15. ———, "Economies as an Antitrust Defense: Correction and Reply," *Amer. Econ. Rev.* Dec. 1968, *58*, 1372–76.
16. ———, "Allocative Efficiency and the Limits of Antitrust," *Amer. Econ. Rev., Proc.*, May 1969, *59*, 105–18.
17. J. C. G. WRIGHT, "Products and Welfare," *Australian Economic Papers*, Dec. 1969.

Teaching Hour-Load Assignments in Economics Departments in Larger Institutions Revisited

By RODNEY J. MORRISON*

The structure of economists' salaries and income has been examined in the studies published recently by the American Economic Association, [2], [3]. Another aspect bearing upon the economic status of the profession, at least for teaching economists, is teaching loads. In 1956, John Due published in this *Review*, the results of a December 1955 survey of teaching loads in the economics departments of twenty-five large institutions [1]. The departments he surveyed were: Northwestern, Michigan, Michigan State, Illinois, Ohio State, Wisconsin, Minnesota, Indiana, Purdue, Yale, Harvard, Princeton, Columbia College, Columbia University Graduate Faculty, Cornell, Pennsylvania, New York University, Carnegie Institute of Technology, Duke, Virginia, Texas, Louisiana State University, Stanford, the University of California at Berkeley, and at Los Angeles.

These same twenty-five institutions were surveyed in February of 1969 by the present author. Twenty-three departments responded, and the results of this survey are presented in Table 1. To determine typical rather than nominal teaching loads, replies to the present study were subjected to some interpretation. This evaluation was minimal, however, as most responses were quite clear-cut. In the present survey no distinction was drawn between graduate and undergraduate teaching, and the results represent the average hour-load per week per academic year. Also, despite interpretation, the results still may be somewhat high, as several departments indicated that in special circumstances "arrangements" could be made.

The median hour-load over all ranks in Due's 1955 study was nine hours, and based upon the data in Table 1, the median load in the present study (1968–69) is but six hours. To facilitate further comparison, Table 2 presents median hourly teaching loads by individual rank for Due's earlier study and median load by individual rank for approximately the same group of schools for the academic year of 1968–69.

From the information contained in these surveys, it appears that the median teaching

* Assistant professor of economics, Wellesley College.

TABLE 1—TYPICAL TEACHING LOADS IN TWENTY-THREE LARGE INSTITUTIONS, 1968–69

Typical Weekly Class-Hour-Load Teaching Assignment per Full Time Equivalent	Number of Institutions Assigning the Various Hour-Load Figures, by Faculty Rank			
	Professor	Associate Professor	Assistant Professor	Instructor
4	2	1	1	1
4.5	1	0	0	0
5	4	5	4	3
6	12	11	9	4
7	1	1	1	1
7.5	1	3	4	0
8	0	0	0	0
9	2	2	4	2
10	0	0	0	0
11	0	0	0	0
12	0	0	0	2

load over all ranks in these arbitrarily selected departments has decreased by a third, falling from nine hours in 1955–56 to six hours in 1968–69. Additionally, it appears that the practice of discriminating by rank has also declined. In Due's study, the median load by rank was nine hours for full and associate professors, ten for assistants, and twelve for instructors. In the 1968–69 survey, the median load was six hours, regardless of rank. Thus assistant professors and instructors may no longer be bearing disproportionate loads. It also appears that there has been a decline in use of the rank of instructor per se, for only thirteen departments reported that they presently staff instructors, while in Due's earlier study twenty-four of the twenty-five departments surveyed used this rank. Further evidence of this change is the fact that several departments that currently do not staff instructors stated quite clearly that the rank was no longer applicable.

To gain some knowledge of the differences in teaching loads between the economics departments at these larger institutions and those of some smaller colleges, the same questionnaire was distributed to twenty-five arbitrarily selected schools. The smaller colleges surveyed were: Coe, Dartmouth, Amherst, Williams, Mt. Holyoke, Smith, Wellesley, Oberlin, Antioch, Carleton, St. Olaf's, Reed, Middlebury, Bowdoin, Trinity (Hartford, Conn.), Swarthmore, Ripon, Miami of Ohio, Grinnell, Claremont Men's College, Wheaton, Haverford, Bryn Mawr, Wesleyan, and Vassar. Twenty-three of these departments responded, and the results are presented in Table 3.

The median hour-load over all ranks in these departments is nine hours. Moreover, the median load for each individual rank is also nine hours, indicating that little discrimination by rank exists. In this respect, these smaller departments resemble the large departments of Table 1. While they may assign a heavier median load over all ranks, it is evident that a load of between six and seven and one-half hours has been adopted by a good number of these smaller institutions as well.

Table 2—Median Weekly Class-Hour-Load Teaching Assignment, by Rank

Rank	Median Load in Hours, 1955–56	Median Load in Hours, 1968–69
Professor	9	6
Associate Professor	9	6
Assistant Professor	10	6
Instructor	12	6

Table 3—Typical Teaching Loads in Twenty-Three Smaller Institutions, 1968–69

Typical Weekly Class-Hour-Load Teaching Assignment per Full Time Equivalent	Number of Institutions Assigning the Various Hour-Load Figures, by Faculty Rank			
	Professor	Associate Professor	Assistant Professor	Instructor
4	0	0	0	0
5	0	0	0	0
5.5	1	1	1	1
6	4	4	4	2
7	1	1	1	0
7.5	4	4	3	2
8	1	1	1	1
9	9	9	10	9
10	1	1	1	1
11	0	0	0	0
12	2	2	2	3

REFERENCES

1. J. F. DUE, "Teaching Hour-Load Assignments in Economics Departments in Larger Institutions," *Amer. Econ. Rev.*, Dec. 1956, *46*, 970–71.
2. N. A. TOLLES AND E. MELICHAR, "Studies of the Structure of Economists' Salaries and Income," *Amer. Econ. Rev., Supplement*, Dec. 1968, *58*.
3. COMMITTEE ON THE NATIONAL SCIENCE FOUNDATION REPORT on the ECONOMICS PROFESSION, "The Structure of Economists' Employment and Salaries, 1964" *Amer. Econ. Rev., Supplement*, Dec. 1965, *55*.

The Restrictive Effect of the U.S. Tariff: Comment

By FRANKLIN V. WALKER*

In a recent issue of this *Review*, Georgio Basevi [2] found that the U.S. tariff structure so improved the terms of trade as to yield an estimated welfare gain of $258 to $558 million in 1958–62. This indicates that unilateral tariff removal by this country would reduce home welfare. In this comment, Basevi's model is modified to estimate the gain to the United States from tariff elimination simultaneously by all countries. In addition, following a clue given by Basevi but not carried out in his own calculations, that part of the gain attributable to change in the net barter terms of trade is separately measured.[1]

I. *The Terms of Trade Effect*

Basevi's model and the present modification of it incorporate the welfare effect of changes in terms of trade which follow removal of tariffs. These price changes occur both because international supply schedules are less than perfectly elastic and because revaluation of exchange rates is in general required to maintain equilibrium in the balance of trade. An elasticity approach model is justified by the assumption that real incomes are kept constant at home and abroad (except for the gain or loss measured by the model) through appropriate governmental policies.[2]

The rationale of the model may be understood by examining Figure 1, which is labelled in conformity with Basevi's figures. In the left-hand panel, S_m is the import supply schedule in dollars facing the United States under protection. The U.S. tariff is TB and the price to consumers is TC. The supply schedule inclusive of tariff is not drawn. The volume of imports is OC and their value, net of tariff, is represented by the area $OABC$. In the right-hand panel, the dollar price of exports under protection is $O'A'$. Units of measurement of exports and imports are so chosen that the export price is equal to the tariff import price, OA, under protection to facilitate comparison of the two panels. The quantity of exports is $O'C'$, which is greater than the quantity of imports, OC, and the value of exports in dollars is $O'A'B'C'$. Thus, under protection the balance of trade is $O'A'B'C'$ minus $OABC$, which is equal to $K'N'B'C'$.

Removal of the U.S. tariff brings an increase in the quantity and value of imports. Simultaneous removal of tariffs abroad expands demand for U.S. exports and their dollar value also rises. A fortuitous set of elasticities would leave the balance of trade unaltered, but in general this case cannot be assumed and some revaluation of the dollar must occur to preserve equilibrium in the balance of trade.

In the case shown in Figure 1 there is devaluation of the dollar. This raises the import supply schedule to S_m' and, together with the effect of foreign tariff removal, shifts the demand for exports to D_x'. The amount of devaluation, and hence, shifts of these schedules is constrained by the requirement that the value of exports under free trade, $O'F'G'H'$, exceed the value of imports, $OFGH$, by exactly the same amount as under protection, $O'A'B'C'$ minus $OABC$.

Basevi measures the gain to the United States from tariff removal by the area $CTGH$, the "...increase in import consumers' utility",[3] *minus* $C'B'G'H'$, the cost of additional exports to the community. With W representing the gain from tariff elimination.

$$W = CTGH - C'B'G'H'.$$

* The author is professor of economics at the State University of New York at Albany.

[1] The author is grateful to the referee for suggesting that the terms of trade effect be separately measured.

[2] Cf. Basevi [2, p. 847].

[3] Quoted from [2, p. 843]. See also Basevi's fn. 8, p. 845, for a more complete interpretation of consumers' and producers' gains. Neither Basevi nor the present writer consider possible secondary effects on surpluses from production and exchange of domestically-traded goods, which might augment or reduce the total gain from free trade to U.S. citizens.

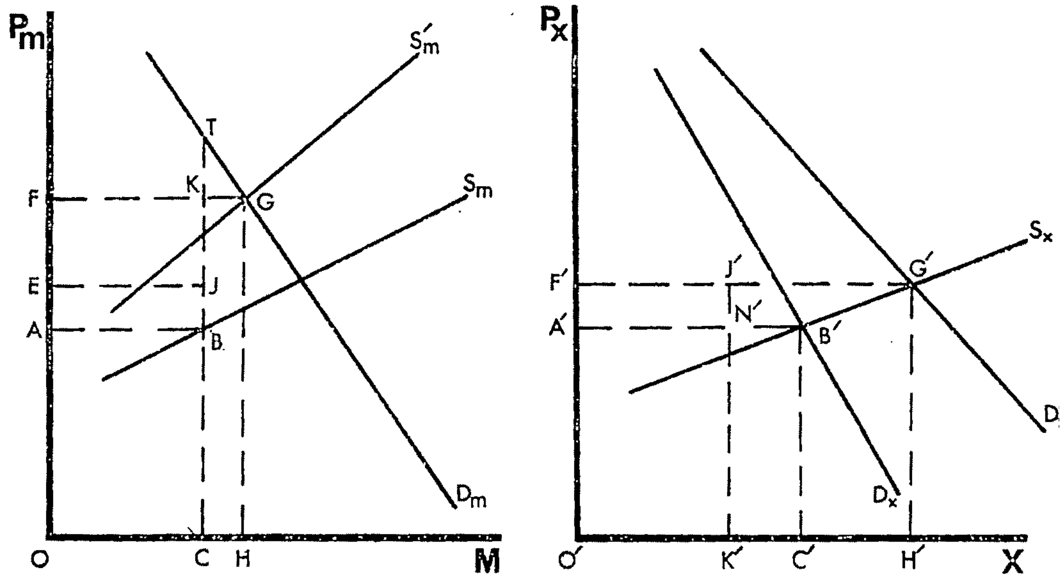


Figure 1

However, $CTGH$ does not represent the net gain to consumers, since an amount represented by $CKGH$ must be paid for the added imports. Consumers' surplus increases by KTG minus $AFKB$, the latter area representing the portion of the tariff paid by foreign exporters by virtue of the lower import price under protection. Home exporters' surplus increases by an amount represented by $A'F'G'B'$. Thus, as Basevi points out, the welfare effect of tariff removal may be measured as

$$W = KTG - AFKB + A'F'G'B'.$$

The gain from trade may arise both from changes in the volume and terms of trade. A terms of trade effect may be obtained by some further rearrangement of the components of W. The differential change in the net barter terms of trade is dp_x minus dp_m, if units of measurement are chosen so that the export price, p_x, and import price, p_m, are both initially unity. In Figure 1, AF represents dp_m while AE, equal to $A'F'$, represents dp_x. Weighting the change in terms of trade by the volume of imports under protection, the terms of trade effect is represented by the area $EFKJ$.

In the right-hand panel, $O'K'$ measures the volume of exports equal to the volum of imports under protection, OC. Thu $A'F'J'N'$ equals $AEJB$. It is apparent the the welfare effect of freeing trade may also b obtained as

$$W = KTG - EFKJ + N'J'G'B'.$$

Algebraically, the terms of trade effec alone, W_t, is

$$W_t = (dp_x - dp_m)M \tag{1}$$

where M is the quantity of imports unde protection. The remaining gain, W_r, :

$$W_r = \tfrac{1}{2}(t_m - dp_m)dM + dp_x(X - M + \tfrac{1}{2}dX) \tag{2}$$

where t_m is the U.S. *ad valorem* tariff rat equal to the tariff since p_m is unity, and X : the quantity of exports. The total gain fro free trade is $W = W_t + W_r$.

II. *Determination of Exchange Revaluatio*

To calculate the necessary revaluation c the dollar, we first set the differential of th balance of trade equation equal to zero:

$$dB = p_x dX + X dp_x - p_m dM - M dp_m = 0 \tag{3}$$

where B is the balance of trade in dollars. The next step is to reduce this to an equation in $\dot{r}$, the proportional change in the exchange rate r.

The differential of export volume may be expressed in terms of export supply and demand slopes as

$$dX = \alpha dp_x = \beta d(\pi_x + T_x) \tag{4}$$

Alpha and *beta* are export supply and demand slopes, respectively. The foreign-currency price of exports before tariff, π_x, equals p_x/r, the dollar price divided by the rate of exchange; units are chosen so that all three of these variables initially equal unity. T_x is the average unit tariff charged abroad on U.S. exports and is the only new variable added to Basevi's model. Since the foreign price before tariff is unity, $T_x = t_x$, the foreign tariff rate on exports.

The change in the foreign currency consumer price with removal of tariffs is

$$\begin{aligned} d(\pi_x + T_x) &= d\pi_x - t_x \\ &= dp_x - \dot{r} - t_x \end{aligned} \tag{5}$$

Substituting (5) into (4) yields

$$dp_x = \frac{\beta}{\beta - \alpha}(\dot{r} + t_x) \tag{6}$$

The differential of imports in terms of slopes is

$$dM = \gamma d(p_m + T_m) = \delta d\pi_m \tag{7}$$

where *gamma* and *delta* are import demand and supply slopes, respectively, and T_m is the U.S. tariff on imports. The latter equals t_m, the U.S. tariff rate, since p_m is initially unity. The foreign-currency price of imports, π_m, is equal to p_m/r. Therefore, $d\pi_m = dp_m - \dot{r}$. With tariff removal, $d(p_m + T_m) = dp_m - t_m$. Thus, solving (7) for dp_m yields

$$dp_m = \frac{\gamma}{\gamma - \delta}(t_m - \dot{r}) + \dot{r} \tag{8}$$

Substituting (4), (6), (7), and (8) into (3) and replacing slopes with corresponding elasticity parameters yields the required proportional rate of change in the exchange rate:

$$\dot{r} = -\frac{\dfrac{X}{M}A_x t_x - A_m t_m}{\dfrac{X}{M}A_x + A_m - 1} \tag{9}$$

where

$$A_i = \frac{\eta_i(1 + \epsilon_i)}{\eta_i - \epsilon_i(1 + t_i)}$$

and $i = x, m$; X/M is the ratio of exports to imports of the United States. *Eta* and *episilon* are demand and supply elasticities, respectively, with their correct algebraic signs.

III. *Estimates*

In estimating the gain from free trade, Basevi's values for all parameters are retained. Elasticities of supply of imports and exports are, respectively, 6.1 and 4.5. Basevi tried two limit values for the elasticity of import demand, −1.5 and −2.7, and two for export demand elasticity, −9.9 and −5.1. The last might be considered too high for a lower limit in light of recent estimates [3]. Consequently, estimates in this note are also made for an export demand elasticity of −2.0. The value of the export demand elasticity plays a more critical role in estimating the gains from free trade than in Basevi's estimates of the gain from the U.S. tariff.

For t_m Basevi employed the U.S. uniform tariff equivalent, derived from effective rates of protection, which he estimated to have been 15 percent in the period 1958–62. Basevi noted this was similar to the 16.7 percent rate for U.S. manufactures in 1962 obtained by Balassa [1] in the same way. Balassa also estimated uniform tariff equivalents of 23.8 percent for the United Kingdom, 17.3 percent for the Common Market, 12.2 percent for Sweden, and 26.4 percent for Japan.

In the calculations of this note, the U.S. tariff rate, t_m, is held at 15 percent. Two values are tried for the rate on exports, t_x,—15 and 18 percent. Values for the volumes of exports, X, and imports, M, are

TABLE 1—ESTIMATES OF THE WELFARE GAIN TO THE UNITED STATES FROM RECIPROCAL TARIFF ELIMINATION
(in $ million for the average 1958–62)

Export Tariff Rate[a] (percent) t_x	Demand Elasticities[b] η_m	η_x	Exchange Appreciation (percent) $-\dot{r}$	Gains from Freeing Trade: Terms of Trade W_t	Other Gain W_r	Total Gain W
	−1.5	−9.9	10.78	$1,358	$588	$1,947
	−1.5	−5.1	9.41	1,189	543	1,731
	−1.5	−2.0	5.08	650	413	1,063
	−2.7	−9.9	7.37	929	772	1,701
15	−2.7	−5.1	5.30	674	674	1,348
	−2.7	−2.0	−0.18	−10	450	440
	−1.5	−9.9	13.79	1,728	696	2,423
	−1.5	−5.1	12.38	1,554	644	2,198
	−1.5	−2.0	7.85	996	494	1,489
18	−2.7	−9.9	9.96	1,245	904	2,148
	−2.7	−5.1	7.72	971	788	1,759
	−2.7	−2.0	1.75	232	523	755

[a] The import tariff rate, t_m, is 15 percent throughout.
[b] The import and export supply elasticities are 6.1 and 4.5, respectively, throughout.

$19,107 million and $15,076 million, respectively, as reported and employed by Basevi. Results of the calculations are shown in Table 1.[4]

IV. *Conclusions*

The gains from freeing trade are estimated to run from about $0.4 billion up to $2.4 billion on the basis of 1958–62 trade figures. Changes in the terms of trade gain dominate and rise as the demand for exports becomes more elastic relative to the demand for imports, as expected. Even at the higher limit, the gain from freeing trade is disappointingly small—$2.4 billion is one-half of one percent of the average Gross National Product in 1958–62. Moreover, since most of the gain arises from changes in the terms of trade it accrues at the expense of welfare in trading partners of the United States and is not a net gain for the world as a whole.[5] Basevi anticipated these results, [2, p. 851], when he wrote "... reciprocal elimination of tariffs, therefore, would involve a net terms-of-trade effect which could be either positive or negative, but still very small."

REFERENCES

1. B. BALASSA, "Tariff Protection in Industrial Countries: An Evaluation," *J. Polit. Econ.*, Dec. 1965, *73*, 573–94.
2. G. BASEVI, "The Restrictive Effect of the U.S. Tariff," *Amer. Econ. Rev.*, Sept. 1968, *58*, 840–52.
3. H. JUNZ AND R. RHOMBERG, "Prices and Export Performance of Industrial Countries, 1953–63," *Int. Monet. Fund Staff Pap.*, July 1965, *12*, 224–71.

[4] In making the calculations, the author followed Basevi's practice of replacing equation (9) with the quadratic equation in $\dot{r}$ obtained when $dp_x dX$ and $dp_m dM$ terms are retained in the balance of payments differential equation (3). The latter, without these terms, was found to underestimate the absolute value of $\dot{r}$ by a maximum of 6 percent of $\dot{r}$ in the calculations for Table 1.

[5] It should be pointed out that these estimates are based on static analysis. They ignore gains from increased competition, a possibly higher growth rate, increasing returns arising from external economies, and other forces that could make the gains from freeing trade far more important than implied by these estimates.

Optimal Policies and Immiserizing Growth

By JAGDISH N. BHAGWATI*

In 1958, I analysed the paradoxical case of "immiserizing growth" [2] where a country, with monopoly power in trade, found that the growth-induced deterioration in its terms of trade implied a sufficiently large loss of welfare to outweigh the primary gain from growth.[1] An obvious corollary of this proposition was that, if the country imposed an optimum tariff (either in both the pre-growth and the post-growth situations, or in the latter situation alone), this paradox would be eliminated.

James Melvin, in an interesting note [5], has now produced yet another analysis of immiserizing growth, where demand differences of the factor-intensity-reversals type are combined with international identity of production possibilities and growth therein to yield immiserization (for one of the two countries) which *cannot* be eliminated by the imposition of an optimum tariff (by the country experiencing immiserizing growth).[2]

This paradox, in a world of paradoxes, is however readily resolved. It can be easily shown that the Bhagwati and Melvin examples of immiserizing growth belong to two quite different *generic* types; and that each, in turn, can be generalized to a different family of immiserizing growth possibilities. The former type can be eliminated by the introduction of optimal policies; the latter cannot.

I. *Distortions and Immiserizing Growth*

The original Bhagwati immiserizing growth phenomenon belongs to the class of cases where a welfare-reducing *distortion* in the economy is the cause of the immiseration. The essential point involved is that the "... gain which would accrue from growth, *if optimal policies were followed*, is outweighed by the incremental loss of real income which the distortion imposes in the post-growth situation vis-à-vis the pre-growth situation" [3, p. 481].

The original Bhagwati example is traditionally analysed in terms of the (primary) gains from growth (at constant terms of trade) being outweighed by the (secondary) loss from worsened terms of trade [2]. A more insightful way of looking at this case, however, is to argue that the assumed free trade and hence failure to impose an optimum tariff (to exploit monopoly power in trade) in both the pre-growth and post-growth situations involves welfare reducing, "distortionary" policies in both these situations. Immiseration then occurs because the gain, which would necessarily accrue from growth if the optimal tariff were imposed in both situations, is smaller than the incremental loss arising from the accentuation (if any) in the post-growth situation of the welfare loss resulting from the distortionary free-trade policy in each situation.

Harry Johnson's [4] further example of immiserizing growth where the country has no monopoly power in trade but imposes a tariff which applies to both the pre-growth and the post-growth situations, is to be explained in terms of the same logic. In the absence of monopoly power in trade, the tariff is necessarily distortionary and, compared with the fully optimal free-trade policy, causes a loss of welfare in each situation. If the growth were to occur with free trade, there would necessarily be an increment in welfare. However, since growth occurs under a tariff, there arises the possibility that the loss from the tariff may be accentuated after growth, and that this incremental loss may outweigh the gain (that would occur under the optimal, free-trade policy), thus resulting in immiseration. Thus,

* The author is professor of economics at the Massachusetts Institute of Technology. Thanks are due to Peter Kenen for helpful comments.

[1] This possibility was also noted and analysed independently by Harry Johnson.

[2] Melvin's geometric analysis uses a case where each country's offer curve, if drawn, would be less than infinitely elastic, thus implying monopoly power in trade for the *other* country.

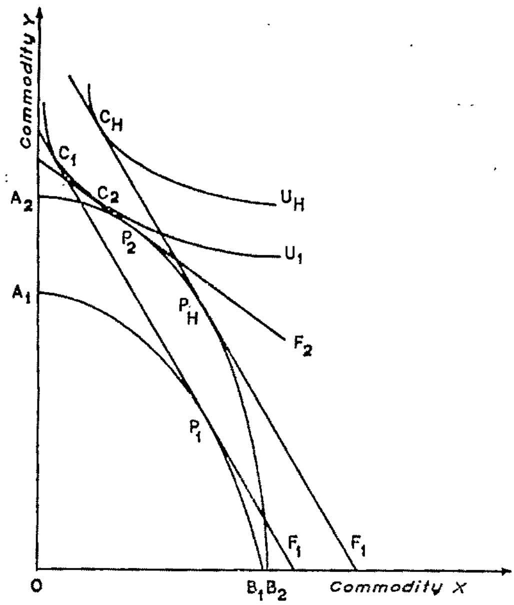


FIGURE 1

the policy-imposed distortion (i.e. the tariff) generates the possibility of immiserizing growth.[3]

I have subsequently produced two more examples of immiserizing growth [3], to underline and illustrate the general proposition that immiserizing growth can arise under *any* kind of distortion, whether endogenous or policy-imposed: first, where the endogenous distortion involves an imperfect factor market, taking the shape of a distortionary wage differential; and second, where the endogenous distortion is foreign, involving monopoly power in trade, and where the country has an optimal tariff in the pre-growth situation.

From the general theory of immiserizing growth, arising from distortions, it is furthermore clear that the removal of a policy-imposed distortion (e.g. removal of the tariff in Johnson's case), or optimal policy intervention to eliminate the welfare-reducing effects of the distortion (e.g. the adoption of a suitable, optimal tariff policy in the original Bhagwati case, or the adoption of a suitable factor-use tax-cum-subsidy policy in the case involving a distortionary wage differential) will eliminate the possibility of immiserizing growth altogether: immiserizing growth can occur only insofar as there is a loss involved in departing from full optimality *and* insofar as such a loss *increases* in the post-growth situation.

II. *Exogenous Reduction in Gains from Trade and Immiserizing Growth*

Consider, however, the case where growth in one country is simultaneously accompanied by growth elsewhere (i.e. the rest of the world). As soon as we change the scenario thus, we are landed with the distinct, and quite unparadoxical, possibility that the primary gain from growth, defined at an *unchanged* foreign offer curve facing one country, and, assuming that optimal policies are followed in the pre-growth and post-growth situations, would be outweighed by the reduction in gain from trade brought about by the *shift* in the foreign offer curve facing the country (resulting from growth abroad). It is clear in such an eventuality that the resulting immiseration accrues really from the fact of an *exogenous* reduction in the gains from trade, obtaining despite the pursuit of optimal policies.

Such a possibility is illustrated most simply with respect to a country with neither monopoly power in trade nor domestic distortions. Remember that, for such a country, free trade is clearly the optimal policy. In Figure 1, the pre-growth situation is depicted by the production possibility curve A_1B_1, the free-trade production and consumption levels by P_1 and C_1, respectively, the given foreign price by F_1, and the welfare level by U_1. If growth were to take place at a *constant* foreign offer (i.e. at constant F_1), then the optimal free-trade policy will imply improvement in welfare at hypothetical level U_H, the post-growth production possibility curve being A_2B_2. However, if we now also assume that the foreign offer curve shifts from F_1 to F_2, production and consumption in the post-growth equilibrium will shift to P_2 and C_2, respectively, and the welfare level will reduce to U_1.[4] The gains from trade will have re-

[3] Augustine Tan [6] has produced an alternative analysis of this case, subsequent to Johnson's paper.

[4] This example would be relevant if the rest-of-the-world was a Ricardian economy and its production pos-

duced to wipe out the primary gain from growth (at an unchanged foreign offer curve). This is the borderline case. If the foreign offer had deteriorated any further, the post-growth equilibrium would have shown immiseration.

This analysis could be readily extended to the case where a country has monopoly power in trade in both the pre-growth and post-growth situations. Immiserizing growth for a country would occur, in this instance, if the primary gain in growth, at *unchanged* foreign offer curve and assuming optimal policies before and after growth, were outweighed by the reduction in the gains from trade resulting from a *shift* in the foreign offer curve.[5] This could be illustrated simply by the use of the Baldwin-envelope technique.

The following observations may then be made about this class of immiserizing growth possibilities.

(1) Since the primary cause of the immiseration is the reduction in the gains from trade resulting from shifts in the foreign offer curve facing the country, even though optimal policies are being followed before and after growth, there is no possibility of devising policies to escape such immiseration. By contrast, the earlier class of immiserizing growth possibilities (Section I) arose precisely from the failure to pursue optimal policies and hence optimal policies could eliminate the immiserizing-growth paradox.

(2) It also follows that any number of such immiseration possibilities could be devised: aside from Melvin's example, we have another case in Figure 1. The required shift in the foreign offer curve may further be brought about by shifts in demand or in production abroad. While, therefore, demand differences play a key role in Melvin's specific example, they are not central to the argument leading to immiserizing growth in this general class of cases.

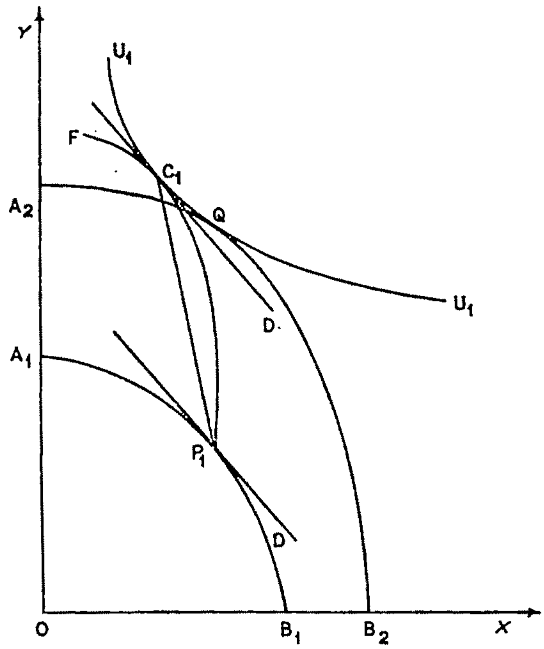


FIGURE 2

(3) The possibility of immiseration would obtain in this class, even if we were to assume that the pre-growth situation was characterised by sub-optimality. In such a case, the reduction in the gains from trade, required to produce immiseration, would merely have to be larger than in the case where the initial pre-growth situation is optimal and hence characterised by a higher welfare level. Melvin's example relates to this kind of comparison: free trade, which is sub-optimal in view of monopoly power in trade, before growth; and an optimal policy after growth. In principle, one could also compare sub-optimal policies both before and after growth and equally well emerge with immiseration.

(4) Finally, note that growth can be immiserizing only insofar as the primary gain from growth is outweighed by the (exogenously) reduced gains from trade. Two corollaries follow immediately. (i) A shift in the foreign offer curve, which increases the gains from trade, cannot lead to immiseration. (ii) The primary gain from growth itself may

sibility curve shifted outwards with bias in favour of the importable commodity. We would further have to assume that our country was operating on the straight line segment of the foreign offer curve from the rest-of-the-world.

[5] The reduction in the gains from trade would have to be measured after the imposition of an optimal tariff relevant to the shifted, foreign offer curve.

be large enough to exceed and dominate the maximum feasible (exogenous) reduction in the gains from trade, thus ruling out immiseration altogether. This is illustrated, for a country with monopoly power in trade, in Figure 2. Before trade, the production possibility curve is A_1B_1. The foreign offer curve facing the country is P_1C_1F and is superimposed on A_1B_1 using Baldwin's technique [1]. Note that the origin of the foreign offer curve now is P_1 (and the offer curve is effectively in the left-hand or fourth quadrant). Figure 2 portrays an optimum tariff policy, with consumption at tariff-inclusive prices at C_1. The resulting welfare level is marked U_1. If then growth were to shift the production possibility curve out to A_2B_2, making it tangential at Q to the pre-growth utility curve, it would be impossible for the country to become worse off than before growth. Indeed, unless gains from trade were to be reduced to zero, the country would become better off than before growth. The possibility of immiserizing growth therefore will arise only insofar as growth is less than that defined by the shift of the production possibility curve to A_2B_2: this is the growth level which implies a primary gain from growth equal to the gains from trade before growth.

References

[1] R. E. Baldwin, "The New Welfare Economics and Gains in International Trade," *Quart. J. Econ.* Feb. 1952, *66*, 91–101.

[2] J. Bhagwati, "Immiserizing Growth: A Geometric Note," *Rev. Econ. Stud.*, June 1958, *25*, 201–5.

[3] ———, "Distortions and Immiserizing Growth: A Generalization," *Rev. Econ. Stud.*, Oct. 1968, *35*, 481–85.

[4] H. G. Johnson, "The Possibility of Income Losses from Increased Efficiency or Factor Accumulation in the Presence of Tariffs," *Econ. J.*, Mar. 1967, *77*, 151–54.

[5] J. R. Melvin, "Demand Conditions and Immiserizing Growth," *Amer. Econ. Rev.*, Sept. 1969, *59*, 604–6.

[6] A. H. Tan, "Immiserizing Tariff-Induced Capital Accumulation and Technical Change," *Malayan Econ. Rev.*, forthcoming.

A Further Test of the Kinky Oligopoly Demand Curve

By JULIAN L. SIMON*

Paul Sweezy's hypothesis that the oligopolist's demand curve has a kink in it [4], is commonly accepted as correct. It is the explanation of oligopoly behavior found in most standard price theory texts, despite the fact that neither Sweezy nor anyone else has shown empirical behavioral evidence in support of the hypothesis.[1] The only relevant empirical work was done by George Stigler [3], the conclusion of which was all against the existence of a kinked curve. Yet the kinky-demand-curve hypothesis continues on its merry way.

Stigler's strongest evidence was a showing that monopolists do not alter their prices more often than do nonmonopolists (oligopolists). His reasoning which makes that evidence relevant is as follows:

> Should monopoly prices be rigid, the forces that explain this rigidity may suffice to explain an equal amount of price rigidity under oligopoly. Unless the factors making for monopoly price rigidity do not operate under oligopoly, we can dispense with the kink unless greater rigidity is found in the oligopolistic industries. [3, p. 420].

Monopoly prices are in fact observed to be far less than perfectly flexible and there is no particular reason to believe that the institutional-organizational forces present in monopolies do not also operate in oligopolies. Hence a kink in the demand curve should indeed result in fewer price changes for oligopolists than for monopolists. That is why Stigler's data that show as many (in fact, more) price changes for nonmonopolists (oligopolists) than for monopolists tend to refute the existence of kinks in demand curves.

Further evidence should be welcome, not only because there is an unresolved tension between Stigler's data and Sweezy's hypothesis, but also because Stigler's sample of monopolists was small—just two firms. This paper therefore tests the hypothesis upon business magazine advertising rates. Business magazines offer a favorable ground for investigation of price behavior because the advertising rates of large numbers of business magazines have been published regularly over a long period of time. Furthermore, the published price is a frequent transaction price, though haggling with major customers sometimes does take place.[2] In general, the magazine industry lives in a statistical fishbowl. But these remarkable and accessible data have not been much exploited.

The data for this study were taken from several editions of *Business Publication*

* The author, professor of economics and of marketing at the University of Illinois, is grateful for the enjoyable experience of presenting this work to A. James Heins' senior honor seminar of 1967–68.

[1] Hall and Hitch's questionnaire study [2] might be considered evidence in support of a kinky curve. In fact, Stigler considers that Hall and Hitch's notion was basically similar to Sweezy's. But their data were impressionistic verbal responses and not actual price changes.

[2] Brown et al. [1, p. 256] say that, unlike television rates, magazine rates "... are largely based upon published schedules ... print deals are infrequent. ..." And the following trade press report provides a view into pricing in the oligopolistic broadcast industry, as well as substantiating that published magazine rates are frequent transaction prices:

> Chicago, April 25—Ward L. Quaal, president of WGN Continental Broadcasting, lashed out last week at "irresponsible" broadcasters who are "using rubber rate cards and fantastic, self-devised interpretations of rating and circulation figures." He advised broadcasters to "put our house in order before the federal government does it for us."
>
> In a talk to the Chicago Advertising Club, Mr. Quaal singled out Los Angeles as a city of extensive rate cutting by radio and tv stations. "In the years 1952 to 1959, when television was enjoying its greatest growth, none of the Los Angeles tv stations showed more than a 10 percent sales gain in any of those years, because they were too busy cutting each other's throats by cutting their advertising rates," he said. ...
>
> The broadcasting industry should follow the lead of other advertising media such as newspapers, magazines and outdoor in gaining the respect of advertisers by "determining what our medium is worth and then sticking by our rates," Mr. Quaal said (*Advertising Age*, May 1, 1967, p. 85).

Circulation and Rate Trends, (*BPCRT*), published by the Association of National Advertisers (*ANA*). Included in *BPCRT* are rate and circulation information on 830 business magazines whose circulation had been audited by the Audit Bureau of Circulations, Business Publications Audit of Circulation, or Verified Audit Circulation Company, and who met the compiler's (*ANA*'s) circulation criteria. These 830 include most of the largest business magazines among the roughly 2,500 published in the United States (as of 1961). The original data were collected by the advertising audit organizations and *Standard Rate and Data Service* (*SRDS*) from their various publications.

The magazines in *BPCRT* include entries in 148 of the 174 classifications established by *SRDS*. From here on, the magazines within each classification category for which *BPCRT* provides information, and with which we shall work, will be called a group. Groups contain anywhere from one to twenty nine publications.[3] The members of a group are not for the most part head-to-head competitors, as are steel firms or tire companies. Rather, the audiences of the business magazines differ from one another to greater or lesser degrees, and hence they are not perfect substitutes; in that respect any magazine's status as an oligopolist is doubtful. And in some cases a publication is clearly a very poor substitute for any other in the group; for example, *American Journal of Cardiology* is classified together with 17 other "Medical and Surgical" publications. And *Midwest Electrical News* is in the same group as national and other regional electrical publications. The basis and meaning of the classification system is discussed later.[4]

The basic data are the year-end rates for each magazine (for a single insertion of a black and white full page advertisement), given for 1961–64 in one edition of *BPCRT*, and in another edition for 1955–61. For each magazine I counted the number of years in which the year-end rate differed from the previous year's. In each case, this must be equal to or less than the total number of rate changes, because the rate may change twice or more within a given year. But such multiple changes should cause no bias and anyway they are rare in this industry. The mean number of change years was then computed for the magazines in each group. (For the single magazine groups, the one magazine constituted the mean.)

It should be noted that the advertising rate is not the true price, because circulation can and does change considerably while the rate remains constant; that is, the effective price which is customarily computed as the

[3] Magazines that were not in existence for the full relevant spans of time were not considered members of a group. Hence a magazine classified here as a "monopolist" for 1955–61, say, could actually have had competition from 1956 to 1961. However, sidechecks indicate that this factor seems only to blur the later results slightly and not to bias them so as to produce specious conclusions.

[4] The following listings of the magazines included in the first four classifications may help orient the reader:

Class 1, Advertising and Marketing: Advertising Age; Advertising and Sales Promotion; Art Direction; Broadcasting; Business Publications Rates and Data; Canadian Media Rates and Data; Consumer Mag. and Farm Pub. Rates and Data; C. A. Journal of Commercial Art and Design; Editor and Publisher; Film and Television; Graphic Arts Buyer; Industrial Marketing; Media Agencies Clients; Media/Scope; Network Rates and Data; Newspaper Rates and Data; Premium Practice; Print Magazine; Printers Ink; Publishers Auxiliary; Signs of the Times; Southwest Advertising and Marketing; Sponsor; Spot; Spot Radio Rates and Data; Spot Television Rates and Data; Television Age; U. S. Radio; Western Advertising.

Class 2, Air Conditioning, Plumbing and Heating, Sheet Metal and Ventilation: Actual Specifying Engineer; Air Conditioning, Heating, and Refrigeration News; Air Conditioning and Ventilating; American Artisan; ASHRAE Journal; Contractor; Electric Heat and Air Conditioning; Electric Heating Journal; Fuel Oil and Oil Heat; Fuel Oil News; Heating and Air Conditioning Contractor; Heating, Piping and Air Conditioning; Heating and Plumbing Merchandiser; Journal of Plumbing, Heating and Air Conditioning; Mechanical Contractor; Plumbing-Heating-Air Conditioning Wholesaler; Plumbing-Heating-Cooling Business; Refrigeration and Air Conditioning Business; Ships; Supply House Times.

Class 3, Amusements: Amusement Business.

Class 4, Architecture: AIA Journal; Architectural and Engineering News; Architectural Forum; Architectural Record; Building Construction; Pacific Architect and Builder; Progressive Architecture; Western Architect and Engineer.

TABLE 1

Group Size (1)	1955–61 Number of Groups (2)	1955–61 Grand Mean Number of Change Years (3)	1961–64 Number of Groups (4)	1961–64 Grand Mean Number of Change Years (5)	Sum of Columns 3 and 5 (6)
1	44	2.1	43	.86	2.96
2	29	2.23	36	1.07	3.30
3	16	2.5	20	.98	3.48
4	8	2.28	13	1.12	3.40
5	7	2.43	12	1.56	3.99
6	9	2.40	8	1.10	3.50
7	2	2.5	3	1.33	3.83
8	2	2.31	5	1.33	3.63
9			2	1.83	
10	2	2.95	3	1.13	4.08
11	1	1.81			
12	3	2.81	2	1.25	4.06
13	3	2.1			
14			2	1.14	
15					
16	2	2.25	1	1.0	3.25
17	1	2.53			
18	1	2.05	1	.94	2.99
19			2	1.39	
20			1	1.05	
21			1	1.0	
22					
23					
24					
25	1	2.68			
26					
27	1	2.15			
28			1	1.0	
29			1	.96	

rate per page divided by the number of thousands of readers, and referred to as the rate per thousand, changes while the page rate remains constant.[5] This matter will be considered later. For now we will simply proceed to work with the absolute number of rate change years.

Columns 3, 5 and 6 in Table 1 clearly indicate that monopolistic magazines with no competitors in the same category do not change price more frequently than do firms with competitors, in contradiction to the kinky-demand-curve hypothesis. In fact, for both the 1955–61 and 1961–64 periods, magazines in one magazine groups actually change price less frequently than do magazines in multiple magazine groups (with a single exception in the 1955–61 period), and there seem to be progressively more frequent price changes as one looks from single magazine to at least ten magazine groups (see especially column 6). Such a progression is in-

[5] Consider these data for three exemplary magazines:

Black-and-White Page Rates (Single Insertion), and Prices Per Thousand Circulation

Year	Vend $	Vend Per M	Men's Wear $	Men's Wear Per M	Marine Engineering/Log $	Marine Engineering/Log Per M
1961	450	44.92	735	29.13	515	37.92
1962	450	41.48	735	29.49	515	36.38
1963	500	47.20	815	32.00	515	33.74
1964	500	47.20	815	30.29	595	31.80

teresting, worthy of test as to its existence and of explanation if it is shown to indeed exist. But the important point in refutation of the kinky-oligopoly theory is not that oligopolists (especially the two magazine group) change price more often, but that they do not change price less often. Table 1 shows the latter to be the case.[6] And it should be noted that this main result appears despite the sloppiness of the categories and the nonhomogeneity of the editorial matter of even any two publications.

One might ask if the observed results might also be explained by circulation changes occurring more often in the multi-magazine group than in the single-magazine groups. That is, actual price—price per thousand circulation—might vary more in single magazine groups, even if rate changes are fewer, if circulation changes less often in single-magazine groups. But a check shows that circulation does not vary more in the two magazine groups—the most important (to the analysis) and most populous oligoppolistic group—than in the one magazine groups. For 1955–61, the median percent circulation changes were .171 and .127 for the one and two magazine groups, respectively. One might expect magazines with competitors to have relatively more variation in their circulations, but circulation changes much more slowly than advertising; furthermore, many people read all the magazines in their trade. Therefore, it is not very surprising when these crude data fail to show greater circulation change for the duopolists than the monopolists. This suggests that the data on number of page rate changes are not vitiated by different amounts of variation in circulation.

One may also question whether there is something in the way that the magazines are classified that causes the observed results.[7] This is a possible cause of false results in all studies of industrial orgaization, but one seldom worried about; usually the classification scheme is simply taken as given. To answer the question in this case we must know why the magazines are classified the way they are. The number (and particular members) that end up in a group must depend upon two forces: 1) the principle of classification used by the classifier(s); and 2) the underlying economic reality. Let us consider them in that order.

1. The classification principle used by *SRDS* is pragmatic and empirical rather than analytical, and is similar to the principle used in the Library of Congress book classification scheme. It is intended to be a pattern that will reduce *SRDS* readers' time to a minimum when looking up data for several magazines, and that will suggest by proximity the largest number of useful alternatives when one looks up one or more magazines. For many groups there is little or no question, e.g., medical and surgical publicacations, and dental publications. The tackiest questions probably arise about the decisions whether to make separate categories, rather than which category to put a magazine into, e.g., should osteopathic publications be separate from medical and surgical publications. To some extent these decisions follow the groupings of firms into trade associations, so the *SRDS* classification

[6] Table 1 also shows a lack of flexibility in business magazine advertising rates generally. The mean rate change period is around once every two or three years—inflexible prices by any test, though certainly not unexcelled in other industries. This rigidity of the published (and transaction) rates in the face of per unit real price changes as well as, necessarily, cost changes to the publisher and demand changes over long periods of time, requires explanation. Monopoly power and inelasticity of demand do not constitute a satisfying explanation because magazines just do not have great monopoly power; other magazines and other media such as direct mail are possible substitutes. Nor is don't-rock-the-boat-ism likely to explain much, because no publication alters its rates frequently, even publications whose monopoly positions are well protected.

The explanation of the price rigidity must lie in the institutional factors of (a) long term rate contracts, (b) long term planning based on existing rates, and the desire not to have to revise plans frequently. In other words, publications act as if their customers will more readily accept one big rate jump rather than many small ones, even though the actual price is changing all the time. If publications dealt directly with the people paying the bills rather than with their advertising agents, the picture might be different, because agencies lose by the nuisance of rate changes whereas they do not lose from a higher cost of advertising. In other industries, firms often take the opposite tack, raising the price in many small increments so the change is not so noticeable.

[7] This line of inquiry is owed to the urging of a referee.

scheme is not all *de novo* but partly follows established trade patterns.

2. The economic reality reflected in the classification is the sort of process that Hotelling [5] and others working in this area have tried to capture in the spatial-competition model. The absolute number of competitors will be affected by the size of the market, of course, e.g., medical and surgical publications will exceed osteopathic in number. But the clustering and dispersion will also be affected by the number of competitors, e.g., duopolists are expected to cluster at the same point, whereas new entrants into multi-firm markets may choose the center of the market *or* some place far from the center where there is no competition. Business magazine groups have the additional complication (as compared to the Hotelling spatial market) that the entire market is not isolated from other markets and there is competition at the borders.

Taken together, the above considerations give no reason to doubt that magazines in single member groups correspond to our usual notion of monopoly. (One might further speculate that some of the magazines in the multi-member groups—above ten members, say—might also have steep demand curves, and hence might act more like monopolists than oligopolists. And, in fact, a hint of this latter effect may be seen in Table 1; the greatest amount of price change is to be found in the 10 member group, and the 16 and 18 member groups show markedly less change than the 10 member group. But this effect is not the focus of interest of this note. Rather, we are concerned with the behavior of the monopolist compared to behavior of oligopolists.)

Summary

Like Stigler's, these data on business magazines do not support the kinky oligopoly-demand-curve hypothesis. They give no indication that oligopolists change price less often than do monopolists, as the kinky-demand-curve hypothesis suggests they would.[8] Institutional-organizational rigidities therefore can account for all the rigidity in rates. This does not prove that business men do not imagine a kinky-demand curve, but only that the sum effect of their mental demand curves plus the quickening effects of competition produces no less flexibility than does monopoly.

What these data, together with Stigler's, do prove, I think, is that a deterministic function of any kind is quite inadequate to represent oligopolistic reality. What is needed is a decision tree that portrays the probabilities of the several possible outcomes and the payoffs in revenue and profit as perceived by the oligopolist.

REFERENCES

1. L. O. Brown, R. S. Lessler, and W. M. Weilbacher, *Advertising Media*. New York 1957.
2. R. L. Hall and C. J. Hitch, "Price Theory and Business Behavior," *Oxford Econ. Pap.*, May 1939, *2*, 12–45.
3. G. J. Stigler, "The Kinky Oligopoly Demand Curve and Rigid Prices," *J. Polit. Econ.*, Oct. 1947 *55*, 432–49; reprinted in G. J. Stigler and K. Boulding, eds., *Readings in Price Theory*, Chicago 1952.
4. P. M. Sweezy, "Demand Under Conditions of Oligopoly," *J. Polit. Econ.*, Aug. 1939, *47*, 568–73; reprinted in G. J. Stigler and K. Boulding, eds., *Readings in Price Theory*, Chicago 1952.
5. H. Hotelling, "Stability in Competition," *Econ. J.*, Mar. 1929, *39*, 41–57; reprinted in G. J. Stigler and K. Boulding, eds., *Readings in Price Theory*, Chicago 1952.

[8] Though it is unrelated to the main point of this paper, these data also permit us to examine the behavior of rates when they are cut rather than raised. In most price studies except studies of agricultural prices, most price changes are price rises, and hence most of our knowledge of elasticities is knowledge about the effects of price increases.

One might hypothesize that dollar rate drops by firms with some monopoly power should be bigger proportionally than rate increases, because a seller wants to draw attention to a rate cut and not to a rate increase. But inspection of rate changes for consumer magazines indicates that this is not so in that case. The only 37 rate drops in the data book for 1940–1963 were compared with the rate increases preceding and following. In the 44 possible comparisons (some changes took place at the beginning or end of series) 22 of the decreases were smaller than the adjacent rate increases, and 22 were larger, a perfect stand off. This suggests no difference in the proportional size of dollar rate increases and decreases.

Reciprocal Externalities and Optimal Input and Output Levels

By PHILLIP E. VINCENT*

In two recent articles in this *Review*, James Buchanan and Milton Kafoglis [3] and William Baumol [1] have reexamined the long accepted proposition that activities which exhibit external economies should be expanded in order to achieve a Pareto optimal allocation of resources.[1] Specifically, Buchanan and Kafoglis, employing the case of immunizations, suggest that where two parties each yield external economies to each other, i.e., the externalities are reciprocal, cooperative purchase of the immunizations may lead to a socially optimum position in which *fewer* immunizations are purchased than would be the case if each party acted independently. Furthermore, the outputs received by each party may be either lower or higher than in the case of independent adjustment.

In attacking the previously held proposition, Buchanan and Kafoglis present a numerical counterexample in the form of a game theoretic matrix. While a counterexample is sufficient to prove their point, it does not indicate the general conditions which must be imposed on the loss functions in order to achieve their results.

The following analysis generalizes the Buchanan–Kafoglis numerical examples, to show when fewer rather than more immunizations should be purchased by the two parties and when less, rather than more, output should be produced. In addition, some comments will be made on Baumol's article which extended and reinterpreted some of their results.[2] The main result proved here is that the existence of economies of scale in the production of benefits (outputs) from immunization is a necessary condition for a lower total number of immunizations (inputs) to be purchased at the Pareto optimum, unless rather implausible conditions are placed on the form of the external effects for the examples worked out by Buchanan and Kafoglis and by Baumol. This assumption was not made explicit and appears to have been overlooked in both articles.[3] On the output side, it is proved for one of the original Buchanan and Kafoglis examples, that a reduced output level for both parties at the Pareto optimum requires that one party's receipt of immunizations must be a more-than-perfect substitute for the other's receipt of immunizations. The latter strong substitution relationship was explored earlier by Buchanan and Kafoglis but was not used explicitly in their discussion of the reciprocal externalities case.

Finally, it is suggested that Buchanan and

* The author is assistant professor, University of Southern California. Research for this paper was performed while he was a postdoctoral fellow, Institute of Government and Public Affairs, University of California, Los Angeles. He is especially indebted to Robert M. Coen, Julius Margolis, Werner Z. Hirsch, James M. Buchanan, Morton Marcus, and Donald Shoup for comments on earlier versions and to an anonymous referee for perceptive questions.

[1] One version of this proposition can be found, for example, in Henderson and Quandt [6, pp. 214–16].

[2] The specific examples used by Buchanan and Kafoglis and by Baumol are considered in detail in the Appendix.

[3] Referring to another example of reciprocal externalities, Buchanan and Kafoglis indicate their conclusions in the following statement:

> If a municipal government should cease, forthwith, to provide any collective police protection, individuals would, surely, respond by hiring private policemen, guards, night watchmen. It seems quite likely that, in such a situation, total resource outlay on providing protection to life and property would be greater than under collectivization. The analysis here suggests that this might be true, *even in the absence of scale factors* [emphasis added]. In other words, even if private policemen should be empowered to arrest thieves on property other than that of their employers, and even if police protection is not provided at increasing returns to scale, collectivization might well reduce total outlay on policemen. [3, pp. 412–13].

Kafoglis's original insight that too much input and output may be employed and produced under individual adjustment is indeed correct for another case where economies of scale are not a necessary condition.[4]

I. *A Model of Reciprocal Externalities*

The following model is based on the case of immunizations employed by Buchanan and Kafoglis in order to make the analyses readily comparable. Input will be defined as the amount of immunizations purchased; output will be defined as the reduction in costs due to incurring the illness for which the immunizations are purchased. In this hypothetical case of immunizations, two parties, A and B, each seek to minimize their own total costs associated with a particular disease. They may purchase immunizations which have a cumulative effect within a given accounting period in reducing the chances and resultant costs of becoming ill; these costs are the expected value of missed work days, hospital bills, etc.

Let ${}_AL(a;b)$ be the total costs to A associated with illness when A purchases the level a of immunizations and B purchases the level b of immunizations. The argument b, which is initially assumed not to be subject to A's direct influence, indicates the presence of external effects on A's expected value function from B's consumption level. A similar function may be written for B.

In the present case, this general expression for an externalities relationship for A may be written explicitly as

$${}_AL = pa + {}_AC(a, b)$$

where p is the price of immunizations and ${}_AC(a, b)$ is the expected value of costs due to illness at the levels of immunization, a and b.[5]

In contrast to the present algebraic formulation, the gaming matrix employed by Buchanan and Kafoglis illustrates the process of mutual adjustment which the two parties go through in reaching an equilibrium position when acting noncooperatively. However, this gaming model does not seem particularly valuable for investigating the conditions which must exist at the collective optimum or in comparisons with the conditions which must exist at the noncooperative equilibrium point.[6]

As Baumol [1, pp. 371–2] and others have suggested, externalities may be defined in terms of the effect of one party's activities either on another party's total costs (or welfare) or on the other party's marginal costs. Both types may exist at the same time. An example of a total type of external economy would be, in the current notation, ${}_AC_b<0$ and a marginal type, ${}_AC_{ab}<0$, where

$${}_AC_b = \frac{\partial_A C}{\partial b} \quad \text{and} \quad {}_AC_{ab} = \frac{\partial^2{}_AC}{\partial a\partial b}$$

are the usual subscript conventions for the partial derivatives of the loss functions. The marginal and total types of externalities appear to be directly related to those externalities which Davis and Whinston [5] term nonseparable and separable. There would be no marginal external effects if an externality relationship were separable. In their anal-

[4] After this analysis was substantially completed, it was discovered by the author that Olson and Zeckhauser [7] had previously discussed the papers of Buchanan and Kafoglis and Baumol in the context of the economics of defense. Although they argued that internalization of externalities implies that production will tend to be placed more with the most efficient producer of external benefits which may then lead to Buchanan and Kafoglis's conclusions, they did not delve into the types of technical relationships explored here.

[5] The usually long lasting effects of immunizations and the attendant impact of time make a realistic analysis of the economics of a disease protection program much more complex than the simplifying assumptions permit. Assume here that a new set of immunizations must be purchased during each accounting period and that the decision on how many to purchase is based on expectations about the environment resulting from others' purchase decisions for this period. It is also important to note that income effects are explicitly excluded in this analysis; individual utility functions are not considered. Thus the model's results apply directly only to production or investment processes. The importance of this assumption is discussed further in fn. 12.

[6] The equilibrium positions are assumed to be stable. Buchanan and Kafoglis note that various problems may exist within a strict bargaining context, e.g. the lack of existence of an equilibrium with strategic behavior or multiple equilibria [3, p. 410, fn. 8].

ysis, Buchanan and Kafoglis indicate they assuming that the externalities are nonseparable [3, p. 408] and Baumol emphasized the marginal type [1, p. 371]. It is assumed herein that each party gives marginal external economies to the other at every point along the cost curves which in the present notation is written as

$$ {}_AC_{ab} < 0 \quad \text{and} \quad {}_BC_{ba} < 0. $$

A and B each seek to minimize their own costs. The first minimization process assumes that the two parties act independently as individuals; each adjusts his purchase of immunizations given his expectations of a given disease environment associated with the immunizations purchased by the other party over whom he has no direct influence.

The first-order conditions for a minimization of costs by each party acting individually are:

$$ {}_AL_a = p + {}_AC_a = 0 \quad \text{and} \quad {}_BL_b = p + {}_BC_b = 0. $$

Then, ${}_AC_a(a_{in}, b_{in}) = -p$, and ${}_AC_a(a_{in}, b_{in}) = -p$, where the subscript notation in indicates that the values of the variables are those existing when each party acts on an individual, noncooperative basis.

The second-order conditions are:

$$ {}_AL_{aa} = {}_AC_{aa}(a_{in}, b_{in}) > 0 $$

and

$$ {}_BL_{bb} = {}_BC_{bb}(a_{in}, b_{in}) > 0. $$

This is analogous to the well-known condition that, under this process of individual adjustment, there must exist diseconomies of scale (decreasing returns) in each party's illness cost functions with respect to his own consumption at the cost minimization point, although the sign of the second derivative in this case is naturally the reverse of the usual result for treatment of a profit maximizing firm with revenue and cost functions.

In contrast, joint or cooperative minimization, implying a movement to a Pareto optimum through internalization of the externalities, requires the minimization of the combined cost function,

$$ L = {}_AL + {}_BL = [pa + {}_AC(a; b)] + [pb + {}_BC(b; a)]. $$

The first-order conditions for a minimum in this case are:

$$ L_a = p + {}_AC_a + {}_BC_a = 0 $$
$$ L_b = p + {}_AC_b + {}_BC_b = 0 $$

Then,

$$ {}_AC_a(a_{jt}, b_{jt}) = -p - {}_BC_a(a_{jt}, b_{jt}), $$

and

$$ {}_BC_b(a_{jt}, b_{jt}) = -p - {}_AC_b(a_{jt}, b_{jt}), $$

where the subscript jt indicates the values of the variables at the joint minimization point.[7]

In the original Buchanan and Kafoglis illustrations, A consumes fewer immunizations at the joint minimum than under individual minimization while B consumes more; but in total, they consume fewer immunizations. Hence, assume that $a_{jt} < a_{in}$, $b_{jt} > b_{in}$, but $a_{jt} + b_{jt} < a_{in} + b_{in}$. The proof that the latter conditions require the existence of economies of scale over some range of A's illness costs function entails a comparison between the levels of ${}_AC_a(a_{in}, b_{in})$ and ${}_AC_a(a_{it}, b_{it})$.

If ${}_AC_a(a_{jt}, b_{jt}) < {}_AC_a(a_{in}, b_{in})$, then from the joint and individual first order conditions

$$ -p - {}_BC_a(a_{jt}, b_{jt}) < -p \quad \text{or} \quad {}_BC_a(a_{jt}, b_{jt}) > 0. $$

That is, for A to consume fewer immunizations under joint minimization, the external effect of A's consumption on B's costs must be such as to increase B's total costs, a type of external diseconomy. It is implausible, particularly in the case of immunizations, that the impact of an increase in an activity yields external diseconomies with respect to another activity's total cost curve given that there are external economies with respect to the latter's marginal cost curve,

[7] It is assumed that compensation arrangements can be made such that the two parties each share in the gains accruing from cooperation. Complications introduced by the obvious possibilities of bargaining or when many parties are involved will not be explored here.

the latter condition having been assumed above to hold at all points.[8]

The alternative to this implausible condition entails assuming that

$$_AC_a(a_{jt}, b_{jt}) > {}_AC_a(a_{in}, b_{in}),$$

so that

$$_BC_a(a_{jt}, b_{jt}) < 0.$$

Now we know that $_AC_a(a_{in}, b_{jt}) < {}_AC_a(a_{in}, b_{in})$, since $_AC_{ab} < 0$ for all points and $b_{jt} > b_{in}$. Hence,

$$_AC_a(a_{jt}, b_{jt}) > {}_AC_a(a_{in}, b_{jt}).$$

The latter condition implies that over at least some of the range between $_AC_a(a_{jt},b_{jt})$ and $_AC_a(a_{in}, b_{jt})$, there must exist economies of scale, i.e., that $_AC_{aa} < 0$ for some distance between a_{jt} and a_{in}, holding b constant at b_{jt}.

Similarly, $_AC_a(a_{jt}, b_{jt}) < {}_AC_a(a_{jt}, b_{in})$, since $_AC_{aa} < 0$.

Hence,

$$_AC_a(a_{jt}, b_{in}) > {}_AC_a(a_{in}, b_{in}),$$

and economies of scale must exist somewhere between a_{jt} and a_{in} where b is held constant at b_{in}.[9]

A careful examination of the numbers in the Buchanan–Kafoglis loss function matrix reveals that in part of the region between the individual and joint minimization points, the matrix entries do exhibit economies of scale.[10] That is, the cost to party A at first declines at an increasing rate as the number of immunizations purchased by A increases (given B's level). After A reaches the level of two immunizations, his loss function exhibits decreasing marginal loss reductions with increased immunizations.

Baumol's algebraic demonstration of Buchanan and Kafoglis's point contains in effect the alternative assumption. His explicit welfare functions do in fact contain the result that there is less total resource use at the joint minimization point when each party yields marginal external economies. Yet one party's welfare function is subject to external diseconomies in the total sense at this point, as well as at the individual minimization point.

Although the condition of economies of scale may be plausible in many actual cases, this assumption is not made explicit in Buchanan and Kafoglis's construction of the numerical example. Ordinarily, one might expect the standard assumption of decreasing returns to scale to hold unless specified otherwise. This result indicates that with cooperation use of the input will be reduced by the producer which exhibits economies of scale, a finding contrary to what might have been expected.

With respect to comparative output levels at the individual versus joint minimum points, there is a further assumption implicit in one of the two examples of reciprocal externalities developed by Buchanan and Kafoglis. In their second example [3, pp. 408–11], the total costs associated with illness for each party, $_AC$ and $_BC$, decrease upon moving from the individual to the joint optimum, i.e.,

$$_AC(a_{jt}, b_{jt}) < {}_AC(a_{in}, b_{in}),$$

and

$$_BC(a_{jt}, b_{jt}) < {}_BC(a_{in}, b_{in}).$$

Hence,

$$_AC(a_{jt}, b_{jt}) - {}_AC(a_{in}, b_{in}) < 0.$$

Subtracting and adding $_AC(a_{jt}, b_{in})$ we obtain

$$[_AC(a_{jt}, b_{jt}) - {}_AC(a_{jt}, b_{in})] - [_AC(a_{in}, b_{in}) - {}_AC(a_{jt}, b_{in})] < 0.$$

[8] James Buchanan and Earl Thompson have suggested to the author that specification of a particular type of total external effect of one activity on another would imply nothing about the sign of the marginal effect which is a purely technical matter. Of course, in the present case, the direction of the marginal effect was specified first, and this leads to, assuming continuous cost functions and derivatives, the statement on plausibility of signs in the text. The emphasis herein on the two different types of externality relationships, total and marginal, appears quite consistent with Baumol's detailed discussion of the definition of externality types [1, pp. 371–2] and with the inference to be made from Buchanan and Kafoglis's explicit reference to nonseparable externalities that (in the present usage) both marginal and total external economies are present.

[9] These results should not be interpreted as suggesting there is a necessary interdependence between the level of b and the existence or level of economies of scale in $_AC$.

[10] See the Appendix for an explicit examination of the counter-examples employed by Buchanan and Kafoglis and by Baumol.

The first term on the left is the change in ${}_AC$, holding a constant, in moving from b_{in} to b_{jt} (this term is negative assuming that there are total external economies at all points.) The second term is the change in ${}_AC$, holding b constant, in moving from a_{jt} to a_{in} (which is also negative since additional a always reduces ${}_AC$). From the direction of the inequality, the first term is greater in absolute value than the second.

But it was assumed above that $b_{it}-b_{in}<a_{in}-a_{it}$. This indicates that the impact of a given change in b on ${}_AC$ is greater than the impact of a larger change in a in moving between the joint and individual minimum points. It appears correct to interpret this as a case of B's consumption of inputs being a more-than-perfect substitute for A's own consumption in A's illness costs function, i.e., at some point A would have lower costs if B received one more immunization than if A received the immunization himself. This type of relationship—which is not entirely implausible—and its implications were analyzed in the first part of the Buchanan-Kafoglis article, but its importance to this particular reciprocal externalities example was not indicated.

As to B's cost function, ${}_BC$, it can be shown that it is feasible in the present case, i.e., where $a_{jt}<a_{in}$ and $b_{jt}>b_{in}$, for ${}_BC$ to either rise or fall in moving from the individual or joint minimum without the more-than-perfect substitute relationship.

II. *An Alternative Case*

To this point the analysis of the Buchanan-Kafoglis contribution has been rather critical. However, the following case, which they did not examine, basically supports their original contention.

In all, there are 16 possible case combinations of relative input and output levels in this two-party model of reciprocal externalities, depending on whether the input and cost levels of A and B rise or fall in moving from the individual to the joint optimum. The analyses in Section I can be applied to all those cases where $a_{jt}+b_{jt}<a_{in}+b_{in}$ and either a or b rises while the other falls. In this section, the case where both $a_{it}<a_{in}$ and $b_{jt}<b_{in}$ is examined. Investigation of the inequality relations between ${}_AC_a(a_{in}, b_{in})$ and ${}_AC_a(a_{jt}, b_{jt})$ as in Section I does not, in the present case, lead to the conclusion that economies of scale are a necessary condition for $a_{jt}<a_{in}$ and $b_{jt}<b_{in}$, although it can be shown that economies of scale are a sufficient condition. There is a condition, however, which must hold with respect to the relative impact of marginal externalities compared with the impact of decreasing returns in one's own production of reduced disease costs.

Once more, assume that ${}_AC_a(a_{jt}, b_{jt})>{}_AC_a(a_{in}, b_{in})$ in order to avoid the total external diseconomies condition.

Subtracting ${}_AC_a(a_{in}, b_{in})$ from both sides, we obtain

$$[{}_AC_a(a_{jt}, b_{jt}) - {}_AC_a(a_{in}, b_{jt})] - [{}_AC_a(a_{in}, b_{in}) - {}_AC_a(a_{in}, b_{jt})] > 0.$$

Reasoning analogously to the proof in the latter part of Section I, and assuming that economies of scale do not exist, i.e., that ${}_AC_{aa}>0$ throughout, while ${}_AC_{a\cdot}<0$, the first and second terms are both negative. Hence, the second term, which is the change in ${}_AC_a$ for a movement from b_{jt} to b_{in}, given a_{in} constant, is greater in absolute value than the first. Now if $b_{in}-b_{jt}<a_{in}-a_{jt}$, then the impact of additional b on ${}_AC_a$ is greater than that of additional a in moving between the joint and individual minimum points. If $b_{in}-b_{jt}>a_{in}-a_{jt}$, then the impact of a on ${}_BC_b$ will be greater than that of b by analogous reasoning.

Although there is a similarity, this condition is not the same as the more-than-perfect substitute relationship derived at the end of Section I on the output side. Rather the result suggests that if the (marginal) externalities are sufficiently strong to outweigh the impact of decreasing returns in one party's illness costs function, if is possible that the joint optimum will be such that both parties use fewer inputs than they did at the individual minimum.[11]

[11] It may be possible to interpret some of Baumol's analysis as already having demonstrated this point. See especially his Theorem 2 [1, p. 364].

On the output side, since both a and b are lower at the joint than at the individual minimum, ${}_AC$ and ${}_BC$ are higher at the joint minimum. Thus, it is quite possible that outputs as well as inputs are overextended in the case of reciprocal externalities given the marginal externalities, decreasing returns relationship derived above.

This is a rather striking result and a variety of aspects of the problem were examined for other possible restrictions on the cost functions which might be necessary, e.g., relationships between total costs and the individual and joint optima, restrictions on the second order conditions. None was found.

Thus, the proposition that the combination or coordination of activities exhibiting reciprocal external economies may lead to reduced input as well as output levels at the social (Pareto) optimum does hold without relying on the existence of economies of scale or a more-than-perfect substitute relationship.[13]

[13] The reader who has been following recent developments in the literature on externalities may wish to contrast the above results with those of a recent exchange between Williams [8], [9] and Brainard and Dolbear [2] on a closely related matter. As noted at the outset, the model used specifically excludes income effects. Williams [8] examines public expenditure benefit spillovers between two different communities. Benefit spillovers are essentially a type of external economy and the social welfare functions of the two communities in his diagrams may be reinterpreted as utility functions for two different individuals. The existence of income as well as substitution effects due to relative price changes are important in William's argument that under certain conditions, cooperation between the two communities may lead to reduced use of resources in producing the good which yields benefit spillovers. The latter result is essentially the same as that of Buchanan and Kafoglis for input use, although the Williams model further differs from the model of this paper in that the benefit spillovers are divisible, i.e., receipt of external benefits by nonmembers of a community reduces the benefits received by members of the community.

However, Brainard and Dolbear [2] have pointed out that Williams implicitly allows a redistribution of income (i.e., one party's real income is reduced) to occur between the communities in arriving at his conclusion. Hence the Paretian condition that neither community suffers a real income loss in moving to a new allocation of resources where externalities are internalized is violated in Williams's derivation. In contrast, this condition is satisfied in the above development. Thus the Buchanan-Kafoglis proposition with respect to inputs appears to apply only to inputs in a production process and cannot be extended to the case of final consumption goods that are arguments in utility functions.

Appendix

The Counterexamples of Buchanan and Kafoglis and Baumol

The following presentations are taken directly from Buchanan and Kafoglis and Baumol to demonstrate how their examples in fact conform to the points developed above. For simplicity, Buchanan and Kafoglis will be quoted extensively since their description of the game theoretic model is quite compact. The following quotation refers to Table 1 which is based on Figure 3 in their article [3, pp. 408–10].

> . . . Each of two persons . . . , A and B, is presumed to be able to take any one of five separate actions, as indicated by the five-by-five matrix. . . . Each person can either do nothing, as indicated by the a_0 row and the b_0 column in the matrix, or he may get one, two, three, or four shots.
>
> . . . players will attempt to minimize the payoffs shown in the matrix. In the upper left-hand corner of each square, we measure the expected value of the costs that catching the disease will impose on each of the two persons, A and B. These are determined by the probability of getting the disease, given the level of immunization undertaken by each person, along with the costs of the disease, if it is contracted. These costs, for A, are shown without parentheses; those for B are shown with parentheses.
>
> . . . immunization shots are available at \$10 each. In the lower left-hand corner of each matrix square [in Table 1], the total cost of purchasing shots for each individual are shown, again with those for B placed in parentheses. The expected costs of catching the disease plus the costs of getting the shots must be added to provide a single negative payoff for each individual. This addition is indicated in the lower right-hand corner of each square matrix. This total payoff will provide the basis for individual decision under independent adjustment.

TABLE 1

A \ B	b_0	b_1	b_2	b_3	b_4
a_0	100 *200* (100) 0 100 (0) (100)	60 *120* (50) 0 60 (10) (60)	45 *88* (23) 0 45 (20) (43)	21 *65* (14) E 0 21 (30) (44)	20 *68* (8) 0 20 (40) (48)
a_1	60 *140* (70) 10 70 (0) (70)	45 *110* (45) 10 55 (10) (55)	35 *85* (20) 10 45 (20) (40)	20 *73* (13) 10 30 (30) (43)	19 *76* (7) 10 29 (40) (47)
a_2	40 *110* (50) 20 60 (0) (50)	32 *91* (29) 20 52 (10) (39)	23 *78* (15) M 20 43 (20) (35)	19 *77* (8) 20 39 (30) (38)	18 *84* (6) 20 38 (40) (46)
a_3	29 *99* (40) 30 59 (0) (40)	23 *86* (23) 30 53 (10) (33)	18 *82* (14) 30 48 (20) (34)	17 *83* (6) 30 47 (30) (36)	16 *91* (5) 30 46 (40) (45)
a_4	20 *90* (30) 40 60 (0) (30)	17 *89* (22) 40 57 (10) (32)	16 *89* (13) 40 56 (20) (33)	15 *89* (4) 40 55 (30) (34)	14 *97* (3) 40 54 (40) (43)

> . . . The matrix is specifically constructed so that a position of equilibrium is attained only when each of the two persons gets two shots; this is shown as the (a_2b_2) box, labeled as M. Should the private behavior of the two result in any other position being attained, at least one of the two persons would have some motivation for independently changing or modifying his behavior.

However, if the two parties act cooperatively, they may reduce their total costs from \$78 at position M (see upper right-hand corner) to \$65 at position $E(a_0b_3)$. Total purchases of immunizations by the two parties are reduced from four to three in moving to the latter Pareto optimal position.

Now the results on the necessity of economies of scale when there are marginal external economies proved above are illustrated by examining column b_2 which indicates the reduction in illness costs (upper left-hand corner of each box) as A's consumption of immunizations increases from zero to four given that B consumes two immunizations. These costs for A fall according to the sequence 45, 35, 23, 18, 16. The marginal reductions in costs are, hence, 10, 12, 5, 2; that is, there are initially economies of scale in terms of the impact of A's own consumption of immunizations on A's costs about the point (a_0b_2). This was proved for the general case above, given that there are total external economies as well as marginal external economies.

Since Table 1 is based on the second Buchanan–Kafoglis numerical example where less total input is used and more total output produced, it can also be used to illustrate the

point on the necessity of a more-than-perfect substitute relationship between a and b in A's costs proved at the end of Section I. Movement between (a_2b_2) and (a_0b_2)—i.e. reducing A's consumption of immunizations by 2—increases ${}_AC$ by 22 (45–23). But then a movement from (a_0b_2) to (a_0b_3) reduces ${}_AC$ by 24. A one-unit change in b has a greater impact than a two-unit change in a on ${}_AC$. This more-than-perfect substitute relationship is perhaps better illustrated by noting that at point (a_1b_1), one more unit of b reduces ${}_AC$ by more than one more unit of a.

In contrast, one of Baumol's basic examples actually makes the alternative assumption indicated above, i.e., that if economies of scale are not present, then the externalities of the total type must be external diseconomies.

Baumol initially assumes an economy of three individuals with respective welfare functions ${}_1W$, ${}_2W$, ${}_3W$, and activity levels a_1, a_2, and a_3 [1, pp. 368–9].[18] The functions are defined as:

$$
\begin{aligned}
{}_1W &= 2a_1a_3 - a_1^2 + a_2^2 + 14a_1 + 22a_3 \\
{}_2W &= 5a_1a_2 - a_2^2 + a_3^2 - 14a_1 + 5a_2 \\
{}_3W &= 2a_2a_3 - a_3^2 + a_1^2 - 5a_2 - 22a_3
\end{aligned}
$$

The activity levels are measured for convenience as deviations from their optimal levels and may take on negative values. Independent adjustment by each individual requires that ${}_1W_1 = {}_2W_2 = {}_3W_3 = 0$. Hence,

$$
\begin{aligned}
{}_1W_1 &= 2a_3 - 2a_1 + 14 = 0 \\
{}_2W_2 &= 5a_1 - 2a_2 + 5 = 0 \\
{}_3W_3 &= 2a_2 - 2a_3 - 22 = 0
\end{aligned}
$$

Solution of this system of equations yields the competitive equilibrium values, $a_1 = 1$, $a_2 = 5$, and $a_3 = -6$. Economies of scale are not present in these welfare functions since ${}_1W_{11} = -2$, ${}_2W_{22} = -2$, and ${}_3W_{33} = -2$.

[18] Baumol's analysis deals in terms of welfare functions to be maximized in contrast to the preceding demonstrations which dealt with costs to be minimized. However, the use of the term welfare functions may be somewhat misleading since the possibility of income and substitution effects do not arise in these formulations.

When the three welfare functions are summed, the following social welfare function is obtained.

$$W = W_1 + W_2 + W_3 = 2a_1a_3 + 5a_1a_2 + 2a_2a_3$$

If the social transformation function is assumed to be $a_1 + a_2 + a_3 = 0$, this social welfare function reduces to

$$W = -2a_1^2 + a_1a_2 - 2a_2^2.$$

Then the joint optimum is determined by setting the first derivatives equal to zero.

$$
\begin{aligned}
W_1 &= -4a_1 + a_2 = 0 \\
W_2 &= a_1 - 4a_2 = 0
\end{aligned}
$$

Hence, at this joint or social optimum $a_1 = a_2 = 0$ and the activity levels a_1 and a_2 are both lower than those at the competitive equilibrium.

However, although activity 1 yields marginal external economies to activity 2 at all points $({}_2W_{21} = 5)$, it yields total external diseconomies at the points where $a_2 = 0$ (since ${}_2W_1 = 5a_2 - 14$). This demonstrates that the Baumol model conforms to the results proved.

References

1. W. J. Baumol, "External Economies and Second-Order Optimality Conditions," *Amer. Econ. Rev.*, June 1964, *54*, 358–72.
2. W. C. Brainard and F. Trenery Dolbear, "The Possibility of Oversupply of Local 'Public' Goods: A Critical Note," *J. Polit. Econ.*, *75*, Feb. 1967, 86–90.
3. J. M. Buchanan and M. Z. Kafoglis, "A Note on Public Goods Supply," *Amer. Econ. Rev.*, June 1963, *53*, 403–14.
4. ——— and W. C. Stubblebine, "Externality," *Economica*, Nov. 1962, *27*, 371–84.
5. O. A. Davis and A. Whinston, "Externalities, Welfare, and the Theory of Games," *J. Polit. Econ.*, June 1962, *70*, 241–62.
6. J. M. Henderson and R. E. Quandt, *Microeconomic Theory*, New York 1958.
7. M. Olson, Jr., and R. Zeckhauser, "Collective Goods, Comparative Advantage,

and Alliance Efficiency," in Roland McKean, ed., *Issues in Defense Economics*, New York 1967.

8. A. Williams, "The Optimal Provision of Public Goods in a System of Local Government," *J. Polit., Econ.*, *74*, Feb. 1966, 18–33.

9. ———, "The Possibility of Oversupply of Public Goods: A Rejoinder," *J. Polit. Econ.*, Feb. 1967, *75*, 91–92.

The Economics of Production from Natural Resources: Note

By RONALD G. CUMMINGS AND OSCAR R. BURT*

Vernon Smith, in his recent article in this *Review* [7], attempts to provide "... a unified theory of production from natural resources" encompassing production from exhaustible, as well as replenishable, resources. There is some question, however, as to the general applicability of Smith's model with reference to the optimum rate of production from *exhaustible*[1] resources, and therefore to his description of the optimum rate of investment in these industries. The purpose of this paper then is twofold: first, to discuss the rather limited nature of Smith's results concerning production and investment within the context of exhaustible resources; and secondly, to present a model that retains Smith's emphasis on the interrelationship of capital and resource extraction, but one that conforms to the established theory concerning the economics of exhaustible resources. The model presented in this paper, however, focuses on the individual firm as opposed to the industry model presented by Smith.

I. *A Critique of Smith's Model*

Consider a firm engaged in the exploitation of an exhaustible resource (e.g., a mine). H. Scott Gordon [2] and Anthony Scott [6] have shown that the firm motivated by the desire to maximize the present value of the stream of profits between the present period ($t=0$) and the time period that exhaustion occurs ($t=t^1$) produces at a rate such that

$$M\Pi(t) \equiv MR(t) - MC(t) = \mu e^{rt} \quad \text{for all } t,\ 0 < t \leq t^1 \tag{1}$$

where MR and MC denote marginal revenues and marginal costs at a given time t, and r is the appropriate discount rate for the firm. The constant μ, within the Gordon-Scott framework, turns out to be marginal profits at the initial time period $t=0$. The term μe^{rt} then satisfies Scott's definition of a "user cost," viz., "... the present value of the future profit foregone by a decision to produce a unit of output today" [Scott, 6, p. 42], where "today" is defined as $t=0$.

From the Gordon-Scott results, then, we see that the optimizing firm does not push production to the point where marginal profits are zero, rather, it postpones the current production of an incremental unit if the present value of the profits which the increment can earn at some future date are at least as great as those which can be earned "today." Marginal profits would be zero at any t only in that special case where the present value of future earnings of an increment (the user-cost) is zero.

In Smith's treatment of mining, the behavioral assumption used is that firms equate price and marginal costs [7]. Pure competition justifies price as equivalent to marginal revenue, and the firm is also assumed to neglect the influence that its rate of extraction has upon future reserves of the resource. But this latter assumption is not palatable at all for a mine because typically a single firm has secure tenure in a resource stock. Further, as we have shown above, it has been well established in the literature on the economics of exhaustible resources that under pure competition, price is equated to the sum of periodic marginal costs and a user-cost that reflects the opportunity cost of an increment of the resource used in current production [6]. Therefore, Smith's behavioral equation neglects the user–cost component of costs and implies an overly rapid rate of exploitation.

* The authors are assistant professor of economics and professor of economics and agricultural economics at Montana State University. This article is a contribution from the Montana Agricultural Experiment Station, Journal Series, No. 997.

[1] "Exhaustible resources", as used in this paper, refers to natural resources, the supply of which is not naturally replenished over the time horizon under consideration.

The one case of mining where Smith's behavioral assumption may be applicable is in ground water mining for agricultural production. The number of firms engaged in production is relatively large and pumping is from a common pool so that each firm can logically ignore the impact that his own rate of withdrawal has on future reserves. This is one of the classic illustrations of a "common property" resource.[2]

With reference to investment in an exhaustible resource industry, earlier work has simplified the production process either by amortizing capital expenditures and including them with variable costs, or by taking the physical plant of the firm as given. This approach has its merit in simplicity, but some interesting economic aspects of the problem are lost by not incorporating investment decisions into the model as suggested by Smith.

In Smith's model, he concludes that if recovery of the resource is profitable then capital flows into the resource at a constant rate [7, p. 423]; in particular, at a rate determined by the equation

$$\dot{K} \equiv \frac{dK}{dt} = \delta[px^0 - C(x^0)] > 0 \tag{2}$$

where K is capital, p the selling price of the resource, x^0 the optimum rate of production determined by the *equality* of MR and MC, and $\delta > 0$ is a behaviorial constant for the industry.

It has been suggested elsewhere in the literature that the existence of excess profits, if they are expected to continue for some length of time, may attract resources for the purpose of exploration and the development of new deposits.[3] For this purpose, equation (2) might well serve as a palatable behavioristic assumption. But this seems incomplete. If price and extraction costs associated with the resource are such that profits are quite high, one might well expect existing firms to expand the capacity of operations at existing mineral sites. In this case, investment would be determined within a context including consideration of the optimum time path of recovery as well as expected prices and costs.

What is needed then is an analytical framework in which the optimum rate of production and the optimum rate of addition to capital stock are analyzed within a context that includes consideration of user-costs and other ramifications peculiar to the exhaustion problem.

II. *The Mining Firm*

In this section a set of decision rules are derived concerning the optimum rates of resource recovery and additions to capital stock for the profit-maximizing firm exploiting a mine.

The following notation is used:

State variables:

$x_t \equiv$ the amount of the resource in stock at time t

$y_t =$ the firms' capital stock at time t

Control variables:

$u_t \equiv$ the amount of the resource to be extracted at time t

$v_t \equiv$ additions to capital stock at time t

With reference to the above state and control variables, the following assumptions and definitions are used.

It is assumed that the resource-stock changes (declines) in each period t by an amount equal to resource-use during that period,

$$x_{t+1} = x_t - u_t \tag{3}$$

Similarly, capital stocks change by an amount equal to investment in period t, thus

$$y_{t+1} = y_t + v_t \tag{4}$$

The variable called capital stock is actually viewed as an index of the degree of capital intensity used in production. Such a measure of capital stock for the firm is admittedly arbitrary and nebulous, but it is a useful abstraction. Most of the objections to its use can be easily overcome by introduc-

[2] For a succinct discussion of the "commonality" problem with reference to ground water, see Millman [5].

[3] For an interesting discussion of the ramifications of exploration, see [8].

ing a vector of capital measures specific to the particular industry, and the analysis is changed very little, but the exposition is more tedious.

Further, it is assumed that all variables are non-negative, and that u_t is bounded from above (at a particular t) by a capacity constraint related to the amount of resource in stock and the amount of capital employed by the firm. The relationship is such that larger resource and capital stocks allow for larger rates of production. The constraint situation can be described mathematically by the inequality

$$u_t \leq h(x_t, y_t), \quad t = 0, 1, 2, \cdots \tag{5}$$

In order to impose diminishing returns to either capital investment or holding of the resource stock, we assume that the function $h(x, y)$ is concave; and of course, its first partial derivatives are non-negative. Both the control and state variables are restricted to non-negative values.

Let the firm's profit function be of the form

$$\Pi^t(u_t, v_t, x_t, y_t). \tag{6}$$

Equation (6) states that the firm's profits in any period depend upon the rate of output and additions to capital stock during the period, as well as on the amount of resource and capital stocks available.

We assume that the firm's objective is to maximize the present value of (6) over the time interval t=0 to t=T, where t=0 is the starting point of our analysis, and t=T is the time at which exhaustion of the resource will occur. Discussion of how T is determined is deferred until later. Thus, the firm wishes to maximize

$$\sum_{t=0}^{T} \Pi^t(u_t, v_t, x_t, y_t)/(1 + r)^t, \tag{7}$$

subject to the constraints of (5), where x_0 and y_0 are known constants, i.e., $x_0 = a$ and $y_0 = b$. This optimization problem can be treated within a "Kuhn-Tucker" [9] analytical framework. The mathematical analysis is relegated to the Appendix and only the results stated below.

For notational convenience in what follows, let subscripts on the functions Π^t and h denote the ordered position of the argument for partial derivatives, e.g., $h_1(x_t, y_t) = \partial h/\partial x_t$ and $\Pi_2^t(u_t, v_t, x_t, y_t) = \partial \Pi^t/\partial v_t$.

If the resource is used at a positive level during period t, relation (A-10) (in the Appendix) holds. Multiplication of that equation by $(1+r)^t$ and transposing terms yields

$$\begin{aligned} &\Pi_1^t(u_t, v_t, x_t, y_t) \\ &\quad = \sum_{i=t+1}^{T} \Pi_3^i(u_i, v_i, x_i, y_i)/(1 + r)^{i-t} \\ &\quad\quad + \lambda_t + \sum_{i=t+1}^{T} \lambda_i h_1(x_i, y_i)/(1 + r)^{i-t} \end{aligned} \tag{8}$$

The left-hand side of equation (8) is the increment in profits during period t that is associated with an increment in the rate of use of the resource stock during that period. The first set of terms on the right-hand side of equation (8) is the discounted sum of increments in profits associated with incrementing resource stocks in each period after t. This sum reflects a cost from using stocks in current production, the cost arising from decreased levels of the resource in future periods and associated higher costs of production as the resource is depleted.

The Lagrange multiplier, λ_t, is the current marginal cost in sacrificed profits associated with the binding constraint $u_t = h(x_t, y_t)$. If the constraint is not binding during period t, then λ_t is zero. In general terms, λ_t is the imputed cost in current value (period t) of the inequality constraints of relation (5).

The last set of terms in the right-hand side of equation (8) is the discounted value of an incremental relaxation of the constraint inequalities in future periods (later than t), where the relaxation is effected through an increment to resource stocks in each period. This set of terms is an indirect cost, via the constraints in future periods, of increasing the rate of depletion of the resource in period t.

The entire right-hand side of equation (8) is the "user cost," or opportunity cost, of the resource stock in period t when evaluated at the marginal rate of use. Notice that the

optimal production plan at period t is only dependent on resource and capital stocks at period t, and rates of depletion of the resource and capital accumulation in periods t and later. We have essentially a Markovian criterion function such that the two state variables, resource and capital stocks, completely summarize all the decision processes up to period t. Knowledge of $\{u_i\}$ and $\{v_i\}$, $i<t$, are not needed to derive an optimal production plan for period t, we need only the magnitudes of x_t and y_t. However, optimal rates of production must be determined simultaneously for all periods later than t in order to obtain the optimal rate for period t.

Of course, capital accumulation rates must be derived simultaneously with rates of production of the resource. Equation (A-11) must hold for periods during which capital investment takes place at a positive level. Multiplication of (A-11) by $(1+r)^t$ and transposing terms yields

$$
\begin{aligned}
-\Pi_2^t(u_t, v_t, x_t, y_t) \\
(9) \qquad = \sum_{i=t+1}^{T} \Pi_4^i(u_i, v_i, x_i, y_i)/(1+r)^{i-t} \\
+ \sum_{i=t+1}^{T} h_2(x_i, y_i)\lambda_i/(1+r)^{i-t}
\end{aligned}
$$

The left-hand side of (9) is the current marginal cost of capital investments in period t, while the first set of terms on the right-hand side is the discounted (to period t) marginal value of capital stocks in all future periods. The last set of terms on the right-hand side of (9) is the discounted value of incrementally relaxing all future constraints on resource-use rates by means of incrementing capital stock in each future period.

In summary, relation (9) can be interpreted as the equating of immediate marginal cost of capital investment with the discounted value of its rewards in future periods resulting from its direct effect on the profit function and indirect effect through the constraint set (5). The direct effect on the profit function of future periods could be either positive or negative ($\Pi_4^t \equiv 0$) since v_t is defined as net investment and thus the profit function must account for maintenance and replacement.

Although our discrete time model does not permit us to take an analytical look at the optimal period of exploitation[4], T^*, we can give a cursory discussion of the possibilities. If T^* is finite, then for any $t>T^*$, $u_t=v_t=0$. It is assumed that there is a cost associated with $v_t>0$ and profits are zero for $u_t=0$. Therefore, the above model is complete for any period $T \geq T^*$, and we merely specify T as arbitrarily large.

Difficulty emerges if the optimal period is infinite, which is certainly a possibility. If, after some arbitrarily large but finite time period, discounted profits are a monotonically decreasing function of time, the period T can be extended, but kept finite, such that the magnitude of the criterion function is essentially the same as for $T \rightarrow \infty$. In fact, the special case where the profit function is independent of time *per se* can be covered rigorously by appealing to a general theorem of Bellman on the functional equation of dynamic programming [1, p. 121].

No attempt is made here to present an industry model that utilizes (8) and (9) as behavioral equations of the typical firm. Mineral industries have been dominated during recent periods by technological change, both as it affects costs and derived demand for the resource. Any industry model would be very deficient without cognizance of the major influence of technical change, and the task of developing such a model is beyond us at this time.

APPENDIX

The optimization problem of the mining firm is to

$$
\begin{aligned}
&\text{maximize } \sum_{t=0}^{T} \Pi^t(u_t, v_t, x_t, y_t)/(1+r)^t \\
\text{(A-1)}\quad &\text{subject to} \\
&u_t - h(x_t, y_t) \leq 0 \\
&u_t, v_t, x_t, y_t \geq 0, \quad t = 0, 1, \cdots, T,
\end{aligned}
$$

[4] The terminal time period T^* is viewed as the time period at which exhaustion obtains under an optimal policy vis-à-vis the rate of resource exploitation. T^* is "optimal" then in this sense.

where the variables are $\{u_t\}$ and $\{v_t\}$, and the variables x_t and y_t are implicitly functions of $u_0, u_1, \cdots u_{t-1}$ and $v_0, v_1, \cdots, v_{t-1}$, respectively. The latter relationships follow from equations (3) and (4) in the text and the fact that x_0 and y_0 are given parameters.

Iterating the difference equations (3) and (4) starting with t=1, 2, $\cdots$, we get

$$\text{(A-2)}\qquad \begin{aligned} x_t &= x_0 - \sum_{i=0}^{t-1} u_i \\ y_t &= y_0 + \sum_{i=0}^{t-1} v_i \end{aligned}$$

We form the Lagrangean expression

$$\text{(A-3)}\qquad \begin{aligned} F(u, v, \lambda) &= \\ &\sum_{t=0}^{T} \Pi^t(u_t, v_t, x_t, y_t)/(1+r)^t \\ &- \sum_{t=0}^{T} [u_t - h(x_t, y_t)]\lambda_t/(1+r)^t \end{aligned}$$

where x_t and y_t are the functions in (A-2), and the discount factor $(1+r)^{-t}$ is used in the t-th constraint equation to give the multiplier λ_t a more convenient economic interpretation. Actually, the Lagrange multiplier is $\lambda_t/(1+r)^t$, but we will be focusing on λ_t as an economic measure and the inclusion of the discount factor makes λ_t a measure specific to period t instead of time zero. With u, v, λ, and partial derivatives defined as vectors, necessary conditions for the optimization problem (A-1) are those of Kuhn and Tucker [9]:

$$\text{(A-4)}\qquad \frac{\partial F}{\partial u} \leq 0$$

$$\text{(A-5)}\qquad u\left(\frac{\partial F}{\partial u}\right) = 0$$

$$\text{(A-6)}\qquad \frac{\partial F}{\partial v} \leq 0$$

$$\text{(A-7)}\qquad v\left(\frac{\partial F}{\partial v}\right) = 0$$

$$\text{(A-8)}\qquad \frac{\partial F}{\partial \lambda} \geq 0$$

$$\text{(A-9)}\qquad \lambda\left(\frac{\partial F}{\partial \lambda}\right) = 0$$

For notational convenience in what follows, let subscripts on the functions Π^t and h denote the ordered positions of the argument for partial derivatives, e.g., $h_1(x_t, y_t) = \partial h/\partial x_t$ and $\Pi^t_2(u_t, v_t, x_t, y_t) = \partial\Pi^t/\partial v_t$.

If at some period t, u_t is positive, relation (A-5) implies that

$$\text{(A-10)}\qquad \begin{aligned} &\Pi_1^t(u_t, v_t, x_t, y_t)/(1+r)^t \\ &\quad - \sum_{i=t+1}^{T} \Pi_3^i(u_i, v_i, x_i, y_i)/(1+r)^i \\ &\quad - \lambda_t/(1+r)^t \\ &\quad - \sum_{i=t+1}^{T} h_1(x_i, y_i)\lambda_i/(1+r)^i = 0; \end{aligned}$$

and likewise, (A-7) implies that for v_t positive,

$$\text{(A-11)}\qquad \begin{aligned} &\Pi_2^t(u_t, v_t, x_t, y_t)/(1+r)^t \\ &\quad + \sum_{i=t+1}^{T} \Pi_4^i(u_i, v_i, x_i, y_i)/(1+r)^i \\ &\quad + \sum_{i=t+1}^{T} h_2(x_i, y_i)\lambda_i/(1+r)^i = 0 \end{aligned}$$

If u_t or v_t is zero, the corresponding above equality is a "less than inequality." A result of major economic importance follows from (A-9): should the constraint $u_t \leq h(x_t, y_t)$ not be binding in period t at the optimal magnitudes of the variables, the Lagrange multiplier λ_t is zero for that period, and if λ_t is positive,

$$\text{(A-12)}\qquad u_t = h(x_t, y_t).$$

References

1. R. E. Bellman, *Dynamic Programming.* Princeton 1957.
2. H. S. Gordon, "The Economic Theory of a Common Property Resource: The Fishery," *J. Polit. Econ.*, Apr. 1954, *62*, 124–42.
3. Richard L. Gordon, "A Reinterpretation of the Pure Theory of Exhaustion," *J. Polit. Econ.* June 1967, *75*, 274–86.

4. M. R. HESTENES, *Calculus of Variations and Optimal Control Theory*, New York 1966.
5. J. W. MILLIMAN, "Commonality, the Price System, and Use of Water Supplies," *Southern Econ. J.*, Apr. 1956, *22*, 426–37.
6. A. SCOTT, "The Theory of the Mine Under Conditions of Certainty," in Mason Gaffney, ed., *Extraction Resources and Taxation*, Madison 1967.
7. V. L. SMITH, "Economics of Production from Natural Resources," *Amer. Econ. Rev.*, June 1968, *58*, 409–31.
8. *Federal Tax Treatment of Income from Oil and Gas*, The Brookings Institution, Washington 1963.
9. *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley 1951.

The Stock of Money and Investment in the United States, 1897-1966

By WILLIAM P. GRAMM AND RICHARD H. TIMBERLAKE, JR.*

The last decade has seen an array of empirical studies for testing possible associations between the stock of money and relevant flow variables, such as consumption and income. These studies originated with the comparison by Milton Friedman and David Meiselman between the relative stability of monetary velocity and the autonomous expenditure multiplier; and they include the comments, criticisms, and further investigations the F-M study has provoked [1], [2], [4], [5], [6], [7], [8], [9].

For good reason, the original study by Friedman and Meiselman uses consumption rather than income as the dependent variable in estimating the relative stability of velocity and the multiplier [4, pp. 175–76]. While a persistently high relationship between the stock of money and consumption is found, the Quantity Theory of Money nonetheless hypothesizes more generally that the stock of money is a major factor in the determination of total money expenditures.[1] A brief supplementary test is included in the F-M study using income as the dependent variable. However, no explicit examination of gross private domestic investment as a dependent spending variable is undertaken. Net private domestic investment is included in the F-M definition of autonomous expenditures for testing the stability of the multiplier. Since consumption correlates better with the money stock than does total expenditure-income, Friedman and Meiselman conclude:

> The correlation between money and autonomous expenditures is so low that adding autonomous expenditures to consumption in order to correlate their sum (income) with money approximates adding a random variable to a variable systematically related to *M* [4, p. 211].[2]

In their criticism of the F-M study, DePrano and Mayer note that the problem involved in defining autonomous elements is complex because "even fixed private domestic investment might be endogenous" due to the acceleration principle [2, p. 732]. The time lags assumed in acceleration theory, however, bring DePrano and Mayer to agree with F-M that fixed private domestic investment must be regarded as autonomous.

The purpose of the investigation made here is to test the effect of the money stock (the assumed independent variable) on gross private domestic investment, measured in both nominal and real terms. By analyzing investment as an induced and dependent variable, this study attempts to determine the impact of the stock of money on investment and to assess the realism of the possible acceleration effect of money acting through consumption and income on investment.

Changes in the stock of money, taken as the independent variable, may affect the dollar value of gross private domestic investment in three ways. First, an increase in the stock of money, *celeris paribus*, tends ini-

* The authors are, respectively, assistant professor of economics, Texas A&M University and professor of finance, University of Georgia. They are indebted to the late David McC. Wright for suggesting this topic for investigation and to James Fortson for helpful comments on earlier drafts of this paper. They are also indebted to two graduate students, J. R. Conner and J. W. McFarland, for technical aid and leg work.

[1] The Friedman-Meiselman analogy of ". . . the half miler being required to run the mile race," due to the use of consumption rather than income as the dependent variable in their analysis, is not necessarily valid [4, p. 176]. Since consumption is only a part of the total money expenditures to which the Quantity Theory applies, the miler in the F-M study is being asked to run a race of something less than a mile.

[2] However, in assessing possible routes for money in influencing spending, F-M state: "An initial increase in the stock of money will tend, *via* its impact on the demand for other assets, to produce an increase in money expenditures on all varieties of them and hence on both what is termed investment and what is termed consumption" [4, p. 220].

tially to lower the market rate of interest. A lower interest rate may induce more investment spending because it allows investors to move out farther along their marginal efficiency of investment schedules. This effect may be designated as the interest-rate effect.

Second, a neutral increase in the stock of money tends to produce a real balance effect in the business sector. One of the channels by which businessmen may attempt to disinvest themselves of cash balances is by expenditures on investment goods. This possible reaction may be labeled the business sector real-balance effect.

Third, an increase in the stock of money, through the real-balance effect in the consumer sector, increases consumer spending. A change in consumer spending, given various production capacity constraints, may generate an inducement to further investment spending—the familiar acceleration effect.

At less than full employment, all of the resultant increases in investment almost certainly would include both nominal and real elements. At full employment, any real increase would have to be at the expense of real consumption (forced savings). If only nominal increases in investment were recorded, no utility would be gained in distinguishing the functional causes of investment from any other kind of spending. That is, the economy would simply experience a general price level effect.

To test for these possible effects, the raw data for narrow money (M_1), defined as the sum of currency and demand deposits outside the banks and government, for wide money M_2, defined as M_1 plus time deposits in commercial banks, and for gross private domestic investment I are serialized annually and first differences taken. The data are then correlated for the various time periods shown in Table 1, a categorization similar to the ones used in other studies of this type. Correlation tests are run for first differences on both nominal n and real values r, the latter being nominal values adjusted by a price index.

The results of the tests for first differences are shown in Table 1 under r_{M_1I} and r_{M_2I}. The most striking feature of the nominal first difference correlation values recorded here is their close similarity to the first difference correlations for the stock of money against income [9]. Correlation coefficients in both cases are approximately .80 for the periods of the 1920's and 1930's. Again, during the 15-year period 1938–53, values obtained both for money against investment and for money against income are not significantly different from zero. This similarity substantiates the credibility gap in recorded spending during the war years. In the late 1950's and early 1960's, first difference coefficients in both cases show recovery toward previous highs with values between .60 and .66. Only in the earlier decades does any significant difference show up.

The reliability of this test is substantiated further by the contrasts of nominal n and real r correlations. When employment is continuously at a high level, real investment can increase at the same rate as money only if the rate of increase in money happens to correspond to the actual rate of investment or produces forced savings. Investment can increase rapidly, with concomitantly rapid increases in money, only when business depression has been pervasive. Under either of these circumstances, correlations for nominal and real values should be close together, and this effect is seen here during the 1930's and 1950's.

When the stock of money increases rapidly at a time of relatively full employment—exemplified in this study by the 1913–20 period[3]—increases in real investment being limited to the secular growth in resources cannot be expected to keep up the pace. However, nominal expenditures should increase due to the price-level effect. Under such conditions, therefore, a disparity between the nominal and real values could be expected. The test results confirm this expectation both for the 1908–21 period and for the 1913–20 period.[4]

[3] The 1938–53 and 1939–48 periods would provide similar test cases if the data for constructing the tests were reliable. Because these periods very probably contain spurious data, nothing can be gleaned from them. They neither prove nor disprove anything.

[4] In the 1920–29 period real values for M_1 and M_2 are markedly different. This difference is the result of a −.597 first difference correlation between changes in

TABLE 1—SIMPLE CORRELATIONS OF FIRST DIFFERENCES IN TWO STOCKS OF MONEY ON FIRST DIFFERENCES OF INVESTMENT AND ON THE INTEREST RATE FOR BOTH NOMINAL AND REAL VALUES, SIMPLE CORRELATIONS OF FIRST DIFFERENCES OF OTHER SPENDING AND OF THE RATE OF INTEREST ON INVESTMENT IN NOMINAL AND REAL VALUES, 1897–1966

	r_{M_1I}		r_{M_2I}		r_{M_1R}[c]		r_{M_2R}		r_{SI}[d]		r_{RI}	
	n[a]	r[b]	n	r	n	r	n	r	n	r	n	r
1897–66	.363	.166	.388	.161	.237	.006	.222	.005	.170	−.247	.327	.265
1897–66 Exc. War Years	.536	.362	.565	.350	.168	.071	.180	.065	.593	.349	.362	.328
1897–08	.559	.398	.467	.265	.665	.538	.644	.510	−.211	−.300	654	.629
1903–13	.590	.441	.652	.485	.545	.338	.521	.313	.366	.234	.779	.762
1908–21	.673	.097	.720	.032	.429	−.306	.409	−.598	.434	−.334	.409	.339
1913–20	.401	.153	.686	.307	.541	−.342	.543	−.561	−.066	−.565	.216	.066
1920–29	.813	.810	.596	−.003	.273	−.428	.243	−.677	.332	−.205	.047	−.030
1921–33	.733	.512	.560	.270	.238	−.198	.197	−.074	.547	.413	215	.152
1929–39	.866	.722	.731	.491	.318	.274	.245	.172	.809	.709	.452	.374
1933–38	.873	.686	.840	.483	−.076	.010	−.241	−.267	.710	.385	.040	.015
1938–53	−.021	−.057	−.101	−.113	−.225	−.558	−.207	−.574	−.554	−.728	.035	−.120
1939–48	−.225	−.146	−.277	−.170	−.617	−.763	−.569	−.747	−.863	−.865	.224	.030
1948–57	.445	.382	.085	.063	.314	−.096	.138	−.377	.025	−.078	.172	.103
1958–66	.615	.593	.418	.338	.682	.634	.302	.217	.533	.572	.823	.799
1929–66	.313	.144	.336	.139	.275	.029	.369	.124	.108	−.285	.443	.340

Sources: Original data for the money supply M_1 to 1957 were taken from: U. S. BUREAU OF THE CENSUS, *Historical Statistics of the United States, Colonial Times to 1957*, Washington, 1960, Series X, p. 646. Data for the period 1958–66 were taken from current issues of the *Survey of Current Business* and the *Federal Reserve Bulletin*. Data for money stock M_2 to 1958 were taken from the Friedman-Meiselman study, [3, pp. 259–60]. Data for the period 1959–66 were taken from current issues of the *Survey of Current Business*. Gross Private Domestic Investment (I) and *Gross National Product* from 1929–50 were taken from: A Supplement to the *Survey of Current Business*, *National Income, 1951* ed., Washington 1951, p. 150. Data for the period 1951–66 were obtained from current issues of the *Survey of Current Business* and the *Federal Reserve Bulletin*. The data for Gross Private Domestic Investment for 1897–1928 were constructed from data found in *A Study of Savings in the United States* by R. W. Goldsmith, Princeton 1955. Other spending S was computed by subtracting Gross Private Domestic Investment from Gross National Product. The interest rate R on 4–6 month prime commercial paper for the period 1897–1957 was taken from the *Historical Statistics of the United States, Colonial Times to 1957*, Washington 1960, p. 654. Data for the 1958–66 period were taken from current issues of the *Federal Reserve Bulletin*. The price deflator was the consumer price index (1954 = 100).

[a] Represents nominal values.

[b] Represents real values.

[c] R is the interest rate.

[d] S is other spending.

The remarkable similarity of money's effects on investment spending and on total spending for so many years may be interpreted in the several ways hypothesized above. First, changes in money could effect changes in investment, which in turn might

real time deposits in commercial banks and changes in real investment.

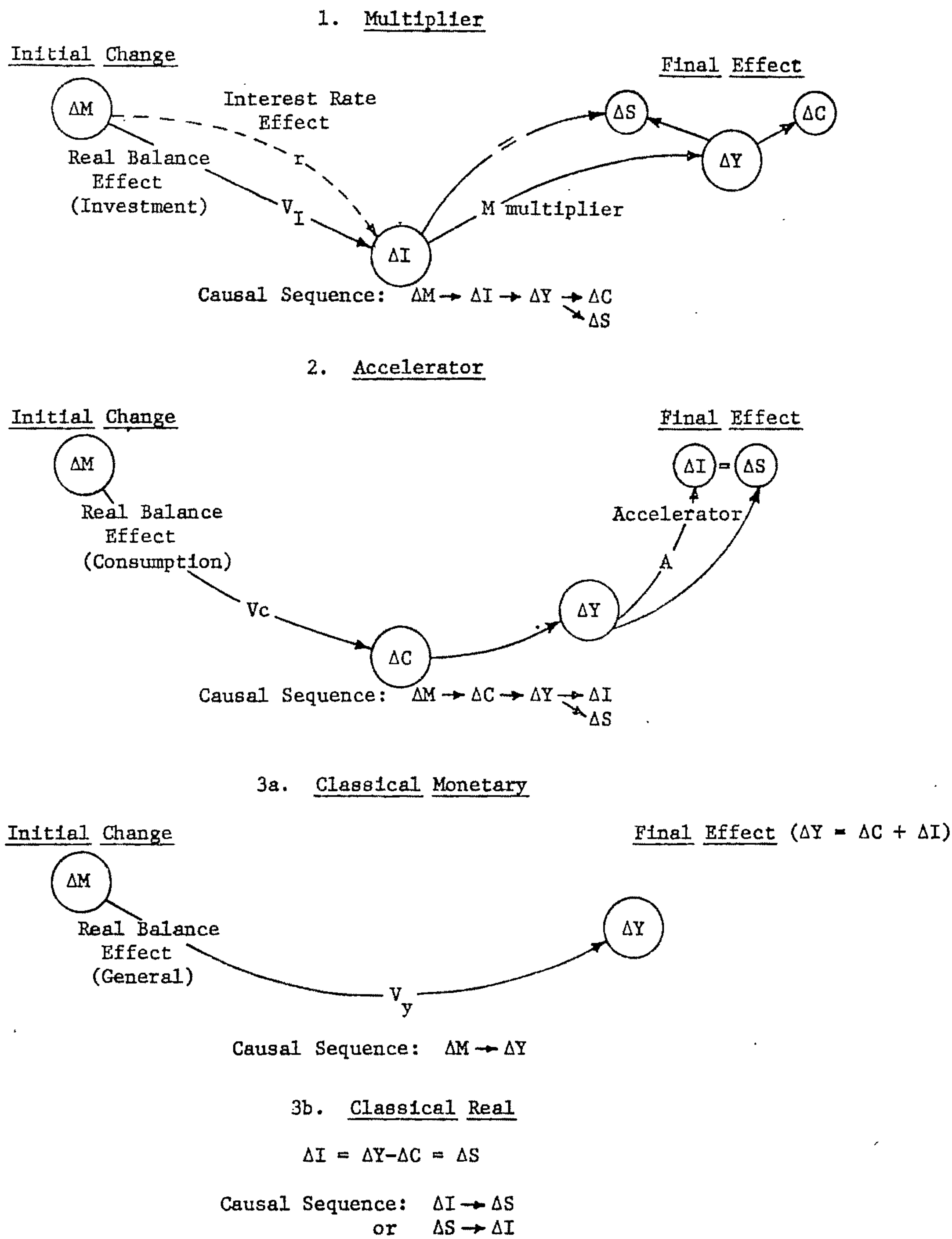


FIGURE 1. DIAGRAM OF POSSIBLE MONETARY EFFECTS

provoke changes in income through a multiplier. Second, changes in money might change consumption that in turn might effect changes in investment through an accelerator. Third, changes in money might prescribe both types of spending simultaneously. Fourth, measured investment spending and measured consumption spending may be in truth taxonomic identities drawn from the same homogeneized population, in which case if money associates well with one it would also correlate well with the other. Certainly, much of what is included in consumer expenditures, such as most consumer durables, is technically investment spending. A similar case may be made for much that is measured as investment expenditure in gross private domestic investment, due to the interdependence between consumption spending and depreciation and inventory changes.

If income is to provoke an accelerator effect, or if investment is to generate a multiplier, consumption spending should correlate positively with investment but with contrasting leads and lags. If classical behavior is hypothesized, an increase in money should be expected to correlate with all nominal spending simultaneously. In real terms at full employment, changes in investment should correlate negatively with other spending (consumption).[5] (See Figure 1.)

In one of the tests conducted here, changes in investment were staged to lag changes in the money stock by one year in order to test for hypothesized accelerator or multiplier effects. In every period this operation made positive correlation values smaller and negative correlation values larger.[6] The same result was experienced in testing for the effect of money on income. These findings suggest a lag of less than a year in money's effect both on what is labeled consumption and on what is labeled investment.

The positive relationship between first differences in nominal money stocks and interest rates that shows up for all time periods except the 1930's and 1940's would seem to deny the generality of a Keynesian liquidity preference thesis. The positive values suggest that any presumed negative effect of changes in the quantity of money on interest rates is ephemeral, i.e., less than a year. As Friedman has pointed out in a recent paper, an increase in the stock of money increases spending and prices and shifts the liquidity preference schedule to the right. When the money stock continues to increase pervasively and at more than a nominal rate, reasonable expectations of further price level increases give additional and positive impetus to interest rates [3, pp. 6–8]. The results of the test here support this contention.

While simple correlations between other spending and investment are consistently high in all periods, the simple correlations of first differences, r_{SI}, are highly variable from period to period (see Table 1). In other tests run, in which investment was lagged one year behind other spending, simple correlations of first differences in most periods were lower than those given in Table 1. One possible explanation for the results in Table 1 is the accounting necessity for other real spending to decline when real investment increases. In real terms the acceleration principle can hold only with unemployed resources. A rise in other spending, given capacity constraints, may produce a stimulus for investment; but at static full employment, resources going to other spending units must decline for real investment to rise. The real, simple correlations of first differences for the periods 1929–39 and 1933–38 seem to exhibit the above mentioned properties though the results are by no means consistent or conclusive.

In general, the tests carried out by this investigation tend to deny the presence of a multiplier or accelerator. At the same time, they seem to substantiate the classical position that a change in the money stock affects all private spending without regard to the categories under which such spending may be defined.

[5] To associate investment with income is to introduce spurious correlation. Thus, investment is correlated with other spending S.

[6] The specific results of the lagged tests are not included here.

REFERENCES

1. A. ANDO AND F. MODIGLIANI, "The Relative Stability of Monetary Velocity and the Investment Multiplier," *Amer. Econ. Rev.*, Sept. 1965, *55*, 693–728.
2. M. DEPRANO AND T. MAYER, "Tests of the Relative Importance of Autonomous Expenditures and Money," *Amer. Econ. Rev.*, Sept. 1965, *55*, 729–52.
3. M. FRIEDMAN, "The Role of Monetary Policy," *Amer. Econ. Rev.*, Mar. 1968, *58*, 1–17.
4. ——— AND D. MEISELMAN, "The Relative Stability of Monetary Velocity and the Investment Multiplier in The United States, 1897–1958," Commission on Money and Credit, *Stabilization Policies*, Englewood Cliffs 1963.
5. ———, "Reply to Ando and Modigliani and to DePrano and Mayer," *Amer. Econ. Rev.*, Sept. 1965, *55* 753–85.
6. ———, "Reply to Donald Hester," *Rev. Econ. Statist.* Dec. 1964, *46*, 369–76.
7. D. HESTER, "Keynes and the Quantity Theory. A Comment on the Friedman-Meiselman CMC Paper," *Rev. Econ. Statist. 46*, Dec. 1964, 364–68.
8. R. H. TIMBERLAKE, JR., "The Stock of Money and Money Substitutes," *Southern Econ. J.*, Jan. 1964, *30*, 253–60.
9. ——— AND J. FORTSON, "Time Deposits and the Definition of Money," *Amer. Econ. Rev.*, Mar. 1967, *57*, 190–94.

Distributed Lags and the Effectiveness of Monetary Policy: Note

By E. PHILIP HOWREY*

In a fairly recent paper [5], Donald Tucker has ably demonstrated that the responsiveness of aggregate demand to monetary policy depends not only on the time form of the response of investment expenditure to changes in the interest rate but also on the response of the demand for money to interest rate changes. In particular, it was shown that a long distributed lag in investment response does not preclude the possibility that aggregate demand responds quickly to changes in the money supply. This argument was put forward to counter the prevalent notion that if investment expenditure responds only slowly to changes in the interest rate, the scope for contracyclical monetary management may be quite limited. While Tucker did show that a long distributed product demand lag is consistent with a rapid response of income to changes in monetary policy, the implication of this finding for the effectiveness of monetary policy was not explored.

The purpose of this paper is to consider rather formally the relationship between the speed of adjustment of income to changes in the money supply and the effectiveness or potential effectiveness of contracyclical management of the money supply. This may seem like a minor point for it might appear to be intuitively obvious that the more quickly income responds to variations in monetary policy, the more effective is contracyclical monetary management likely to be, provided, of course, the stabilization policy is designed properly. Yet, it has become increasingly apparent to students of economic dynamics that intuition can be very unreliable when it comes to the analysis of dynamic models. For this reason alone it seems desirable to subject this proposition to formal analysis. Moreover, as will be demonstrated by the analysis that follows, the proposition that monetary policy is more effective the more quickly income responds to changes in monetary policy is *not* true in general.

The relationship between the speed and effectiveness of monetary policy is developed here within the context of a simple stochastic model of income determination. The basic properties of this model are described briefly in Section I. The stabilization problem is explicitly formulated in Section II. It is assumed that the money supply is varied in such a way that optimal feedback control is applied to the system in order to mitigate as quickly as possible the effects of random disturbances. The residual variance, that is, the variance of income about some target level that remains even when the money supply is varied to minimize instability, is derived and expressed as a function of the lag parameters of the system. An examination of this expression reveals that it is not true in general that the residual variance is smaller the more quickly income responds to changes in the money supply. Indeed, within certain limits, just the reverse is true: the residual variance is *larger* the more quickly income responds to changes in the money supply.

I. *A Stochastic Model of Income Determination*

In order to explore the relationship between the speed with which income responds to money supply changes and the (potential) effectiveness of monetary policy, it is necessary to introduce a dynamic model of income determination. The model that will be analyzed here is Tucker's expositional model in

* This research was performed in the Econometric Research Program at Princeton University with support from the National Science Foundation (GS 1840) and the Office of Naval Research (N00014-67 A-0151-0007, Task No. 047-086). While retaining full responsibility for any errors, the author would like to thank W. J. Baumol, R. E. Quandt, L. G. Telser, and D. P. Tucker for their helpful comments on an earlier draft of this paper.

which distributed lags are included in the behavioral relationships. However, rather than retain the deterministic system as it was originally advanced, random disturbances are explictly introduced into the behavioral equations. The resulting stochastic system is somewhat more realistic [2] and provides a more suitable framework within which to discuss stabilization policy [4] than its deterministic counterpart.

The equations of the model are as follows [5, p. 435]:

$$(1) \quad I_t = (1-j)(a_1 + b_1 Y_t + dr_t) + jI_{t-1} + u_{1t}$$

$$(2) \quad C_t = (1-j)(a_2 + b_2 Y_t) + jC_{t-1} + u_{2t}$$

$$(3) \quad Y_t = C_t + I_t$$

$$(4) \quad M_t^* = (1-m)(e + fY_t + gr_t) + mM_{t-1} + v_t$$

$$(5) \quad M_t = M_t^*$$

where

C = consumption expenditure;
I = investment expenditure;
Y = income;
M^* = desired money balances;
M = money supply;
r = interest rate;
t = time subscript;
j and m are lag parameters with $0<j$, $m<1$; u_1, u_2, and v are random disturbances.

Since price and wage changes are ignored, this model is concerned with the impact of monetary policy on income and employment and thus applies to a situation in which income can expand without encountering a labor-force constraint. Both product demand, $(C+I)$, and the demand for money, (M^*), are assumed to be stochastic rather than purely deterministic functions and each of these demands responds with a distributed lag to changes in the interest rate. The equilibrium conditions (3) and (5) require income and the interest rate to adjust each period so that the stochastically given product and money demands are satisfied.

The dynamic characteristics of this system are determined by the simple distributed lags that are included in the demand equations (1), (2), and (4) of the model. The lag parameters j and m have been introduced in such a way that the static and comparative static properties of the system are not affected by changes in these parameters. For example, if the random variable u_{1t} is suppressed for the moment and investment expenditure is set equal to its lagged value, it is clear from (1) that the steady-state value of I does not depend on the lag parameter j. Although the equilibrium response of investment to a change in the interest rate does not depend on the lag parameter j, the speed with which investment approaches its equilibrium value does depend on this parameter. The mean lag implied by the relationship between investment and the interest rate is $j/(1-j)$.[1] This expression for the mean lag indicates that the response of investment to a change in the interest rate is quite rapid if j is close to zero and is much slower if j is close to one.

The final form of the equation for income can be obtained from the system of equations (1)–(5) by solving for Y in terms of M and the disturbance terms:

[1] The partial relationship between investment and the interest rate implicit in equation (1) can be written as

$$(A) \qquad (1 - jL)I_t = (1-j)dr_t$$

or

$$(B) \qquad I_t = (1-j)d(1 + jL + j^2L^2 + \cdots)r_t$$

where L, the lag operator, is defined by $L^k y_t = y_{t-k}$. It is clear from (A) that the equilibrium change in investment corresponding to a unit change in the interest rate is d. From (B) it follows that $(1-j)$ of this change takes place in the same period that the interest rate is changed, $(1-j)\,j$ takes place in one period after the interest rate change and, in general, $(1-j)j^n$ of the change takes place in the nth period after the change in the interest rate. The mean lag of the distribution is therefore

$$(C) \qquad \bar{\theta} = (1-j)\sum_{n=0}^{\infty} nj^n = j/(1-j).$$

In terms of the generating function of the lag distribution, $W(\lambda) = (1-j)d/(1-j\lambda)$, the mean lag is given by

$$(D) \qquad \bar{\theta} = W'(1)/W(1)$$

where $W'(1)$ is the derivative of the generating function evaluated at $\lambda = 1$.

$$Y_t - AY_{t-1} = B + D(M_t - mM_{t-1}) + \epsilon_t \tag{6}$$

where

$$A = \frac{j}{1-(1-j)(b_1+b_2-df/g)} \tag{7}$$

$$B = (1-j)(a_1+a_2-de/g)A/j \tag{8}$$

$$D = \frac{(1-j)dA}{(1-m)gj} \tag{9}$$

$$\epsilon_t = \frac{A}{j}u_t - Dv_t \tag{10}$$

$$u_t = u_{1t} + u_{2t}. \tag{11}$$

It is clear from (6) that the time path of Y will depend on the path of M as well as the values assumed by the random variables u in the product market and v in the money market. It will be assumed that these random variables are mutually and serially uncorrelated with zero means and constant variances denoted by σ_u^2 and σ_v^2.

Within the context of this model, it is now possible to indicate quite precisely how the speed with which income responds to changes in the money supply depends on the lag parameters j and m. Suppose that M_t exhibits a unit increase at $t=1$. The expected income response to this change in the money supply is

$$Y_t' = D(1-m)/(1-A) + A^{t-1}(m-A)/(1-A) \qquad t = 1, 2, \ldots . \tag{12}$$

This is the path that income would follow if there were no disturbances in the system, that is, if $u_t = v_t = 0$ for all t. Now, as Tucker has pointed out, the second term on the right-hand side of (12) depends on the difference $m-A$. If m is equal to A, the response of income to a change in the money supply is complete in the same period that the change in the money supply takes place. And this is true despite the fact that the product demand lag may be quite long (i.e., j and therefore A may be close to unity[2]) provided the money demand lag is correspondingly long. This constitutes Tucker's disproof of the conjecture that a long product demand lag implies that income responds slowly to changes in the money supply.[3]

II. *Income Response and the Effectiveness of Monetary Policy*

We now come to the crucial question of this paper. Is there any reason to believe that stabilization policy is likely to be more effective the more quickly income responds to changes in the money supply? In terms of the model expressed in equations (1)–(5), it would be desirable to know if a small difference between the generalized product demand lag (A) and the money demand lag (m) enhances the (potential) effectiveness of contracyclical monetary policy. One way to answer this question is to modify the basic system to include a plausible contracyclical money supply equation and then to compare the resulting time paths of income for different pairs of values of the lag parameters j and m. It is clear that the time path of income and hence the effectiveness of stabilization policy in general depends upon how well the specific policy that is followed is designed as well as the inherent limitations imposed by the economic system. In order to separate these two aspects of the problem as completely as possible, it is essential to include a control equation that is not determined on an *ad hoc* basis, but rather is designed to effect optimal control.

[2] From the definition of A in (7) it is easy to verify that $A=0$ when $j=0$ and $A=1$ when $j=1$. Moreover, sign $[\partial A/\partial j]$ = sign $[1-b_1-b_2+d\ f/g]$ for all j in the closed interval [0, 1]. Therefore $\partial A/\partial j>0$ and A can be taken as a proxy for j. This justifies the designation "generalized product demand lag parameter" for the parameter A.

[3] This result can also be obtained by calculating the mean lag of the relationship between income and the money supply given in (6). It is not difficult to verify that

$$\theta = W'(1)/W(1) = A/(1-A) - m/(1-m)$$

where $W(\lambda) = D(1-m\lambda)/(1-A\lambda)$ is the generating function of the lag relationship between Y and M. This expression indicates that in this model the mean lag between income and the money supply is the difference between the mean product demand lag $A/(1-A)$ and the mean money demand lag $m/(1-m)$.

Within the context of linear stochastic systems, it seems plausible to consider linear stabilization policy that is designed to minimize the variance of income. Returning to equation (6), the final form of the income equation, it is necessary to determine a control loop on M_t which minimizes the variance of income.[4] In order to derive a minimum variance control equation, it is assumed that the full employment level of income, Y_t^f, is known in advance, but that the money supply for period t must be determined at the end of period t-1 before the actual level of income for period t is known. The minimum variance control equation based on this one period ***inside*** lag of monetary policy is

$$(13) \quad M_t = mM_{t-1} + (Y_t^f - AY_{t-1} - B)/D.$$

It is clear that this is the minimum variance stabilization policy since it reduces the income equation to

$$(14) \qquad Y_t = Y_t^f + \epsilon_t$$

which indicates that actual income differs from full employment income by an amount that depends on the unpredictable disturbances that occur in the current period.

From this last expression, it is apparent that the variance of income with optimum control is simply the variance of the composite disturbance term ϵ_t defined in equation (10). The variance of ϵ_t provides one measure of the potential effectiveness of stabilization policy; the smaller the variance of ϵ_t, the more effective is the stabilization policy. In terms of the parameters of the model specified in (1)–(5), the variance of ϵ_t is given by

$$(15) \qquad \sigma_\epsilon^2 = (A/j)^2\sigma_u^2 + D^2\sigma_v^2$$

[4] Here we abstract from the fact that it is not possible to control the supply of money directly. Since the monetary authority has only indirect control over the amount of money in existence, it might be reasonable to add a disturbance term to this equation to reflect the fact that the desired change in the money supply is imperfectly achieved. This is left as an exercise for the reader.

where σ_u^2 (σ_v^2) is the variance of the disturbance term in the product demand (money demand) equation. By differentiating this variance expression with respect to the lag parameters j and m, it is possible to determine the relationship between the effectiveness of monetary policy and the speed with which income adjusts to changes in the money supply. It is easy to verify that

$$(16) \qquad \partial\sigma_\epsilon^2/\partial m = 2D^2\sigma_v^2/(1-m)$$

which is clearly positive. This shows that the residual variance, that is, the variance of income which remains with the most timely stabilization policy, increases as the money demand lag increases.[5]

This is sufficient to demonstrate that it is not true in general that monetary policy is more effective the more rapidly income responds to changes in the money supply. In order to see this, consider two economies which are characterized by the system of equations (1)–(5). Suppose that the only difference between the two economies is that money demand responds more quickly to changes in the interest rate in one economy that in the other, i.e., $m_1 < m_2$. If A, the generalized product demand lag parameter, is greater than m_2, it follows from equation (12) that in economy two, income responds more quickly to changes in the money supply than it does in economy one.[6] However, we have

[5] This result has been derived on the assumption that the disturbances in the product and money markets are uncorrelated. If the covariance between u and v, σ_{uv}, is not zero, the variance about the full employment level of income is

$$(15') \qquad \sigma_\epsilon^2 = (A/j)^2\sigma_u^2 - 2(DA/j)\sigma_{uv} + D^2\sigma_v^2.$$

the partial derivative of σ_ϵ^2 with respect to m becomes

$$(16') \quad \partial\sigma_\epsilon^2/\partial m = -2DA\sigma_{uv}/j(1-m) + 2D^2\sigma_v^2/(1-m).$$

This shows that if the product and money market disturbances are negatively correlated, the conclusion derived in the text continues to hold. However, if these disturbances are positively correlated, the result in the text could be reversed.

[6] In terms of the mean lag defined in footnote 3, the average lags in the effect of monetary policy for the two economies are

$$\theta_1 = A/(1-A) - m_1/(1-m_1)$$

and

just seen that the residual variance of income is larger when the lag parameter m is greater so that with optimal monetary control the variance of income is larger in economy two than in economy one. Therefore, monetary policy is more effective in economy one than in economy two even though income responds more slowly to changes in the money supply in economy one. This shows that a slow income response can actually increase the effectiveness of monetary policy.

III. *Concluding Remarks*

From this analysis of a linear stochastic model in which investment demand and the demand for money exhibit distributed lag responses to changes in the interest rate, we conclude that it is *not* true in general that monetary policy is likely to be more effective the more rapidly income responds to changes in the money supply. It is entirely possible for a well conceived stabilization scheme to effect a greater reduction in the variance of income about some target level if income responds slowly to money supply changes than if it responds quickly. Thus a rapid response of income to changes in monetary policy is not a necessary condition for stabilization policy to be effective.

$$\theta_2 = A/(1 - A) - m_2/(1 - m_2).$$

With $m_1 < m_2 < A$, it follows that $\theta_1 > \theta_2 > 0$. For example, if $A = .9$, $m_1 = .5$, and $m_2 = .8$, the corresponding average lags are $\theta_1 = 8$ and $\theta_2 = 5$ so that monetary policy works more quickly in the mean lag sense in economy two than in economy one.

References

1. W. J. Baumol, "Pitfalls in Contracyclical Policies: Some Tools and Results," *Rev. Econ. Statist.*, Feb. 1961, *43*, 21–26.
2. G. H. Fisher, "Some Comments on Stochastic Macro-Economic Models," *Amer. Econ. Rev.*, Sept. 1952, *42*, 528–39.
3. M. Friedman, "The Effects of a Full-Employment Policy on Economic Stability: A Formal Analysis," in *Essays in Positive Economics*. Chicago 1953.
4. E. P. Howrey, "Stabilization Policy in Linear Stochastic Systems," *Rev. Econ. Statist.*, Aug. 1967, *49*, 404–11.
5. D. P. Tucker, "Dynamic Income Adjustment to Money-Supply Changes," *Amer. Econ. Rev.*, June 1966, *56*, 433–49.

The Interaction of Tariffs and Quotas

By TSVI OPHIR*

Several recent papers [1], [2], [3], [4] discussed the equivalence of tariffs and quotas. The present note deals with the case of both a tariff and a quota being applied to the same good, and their interaction is explored. In contrast to the earlier papers, it is assumed here that the local producer is a monopolist, whereas imports (though limited by the quota) are traded under competitive conditions. It will be shown that, under such conditions, certain paradoxical results obtain: an increase in the tariff may cause a reduction in price (with increased domestic output), and it may lead to a reduction in domestic output (with a higher price). An increase in the quota may reduce, or even eliminate, actual imports and raise domestic production.

A local monopolist is faced with the competition of imports which are limited by a quota. The supply of imports, up to this limit, is perfectly elastic at the "import price" which includes a tariff. The analysis of the situation is based on the diagram of Figure 1. The market demand curve is $D'D$; the monopolist is faced with a demand curve net of imports, $ABCD$ (where BC equals the quota). The corresponding marginal revenue curve of the monopolist is $AEBCFG$. This MR curve is intersected thrice by many marginal cost curves; two alternative curves, MC_1 and MC_2, are drawn in, with points of intersection H_1, I_1, J and H_2, I_2, K respectively.[1] In such cases there are two local profit maxima. At one of these (at H_1 or H_2, respectively) the quota is effective[2] and fully utilized, with the local producer monopolizing the residual market (with the market price exceeding the import price).[3] At the other (determined by J and K, respectively) the tariff is effective, and the market price equals the import price. There are two subcases. If MC intersects MR in segment BC (e.g., MC_1 at J), there are some imports, and the local producer is in a situation of perfect competition. But if the intersection is in segment CF (MC_2, at K), there are no imports and the local producer exercises some monopoly power, albeit reduced by the *potential* competition of imports.[4]

Which of the two local maxima is the *maximum maximorum*? This can be determined in Figure 1 by a comparison of areas.[5] With $S(EH_1I_1)>S(I_1BJ)$, maximum profit will occur at H_1, and with $S(EH_1I_1)<S(I_1BJ)$, it will occur at J; similarly, for the comparison of $S(EH_2I_2)$ and $S(I_2BCK)$. In Figure 1, both MC_1 and MC_2 are drawn so as to make the compared areas equal, thus indicating the borderline situation of equal maximum profits at H_1 and J (and at H_2 and K). Any decrease in either quota or tariff will move the overall maximum to H_1 (or H_2), and any increase in either of them to J (or K).

From the proceeding analysis, one can directly derive the effects of changes in the quota, the tariff, or production costs on the level of imports, domestic market price and domestic output. These results are presented below in diagrammatic form, with brief comments added.

In Figure 2, the dependent variables are plotted as functions of the quota, Q. With

* Hebrew University, Jerusalem.

[1] Strictly speaking, the MR curve is discontinuous and composed of the three segments; AE, BC, and FG. The loose language of the text is used because it is convenient—and innocuous—to regard points like K as points of intersection of MC and MR.

[2] "Effective" means being the determining factor for imports, price, and domestic output. Obviously, there is always one such factor—*either* the quota *or* the tariff is effective.

[3] Windfall profits accrue to recipients of import licenses.

[4] With constant or falling marginal costs, an intersection as at J does not indicate a local profit maximum —indeed, with falling MC such an intersection corresponds to a profit *minimum*.

[5] S $(ABC\ldots)$ is used below to denote the area of the polygon $ABC\ldots$.

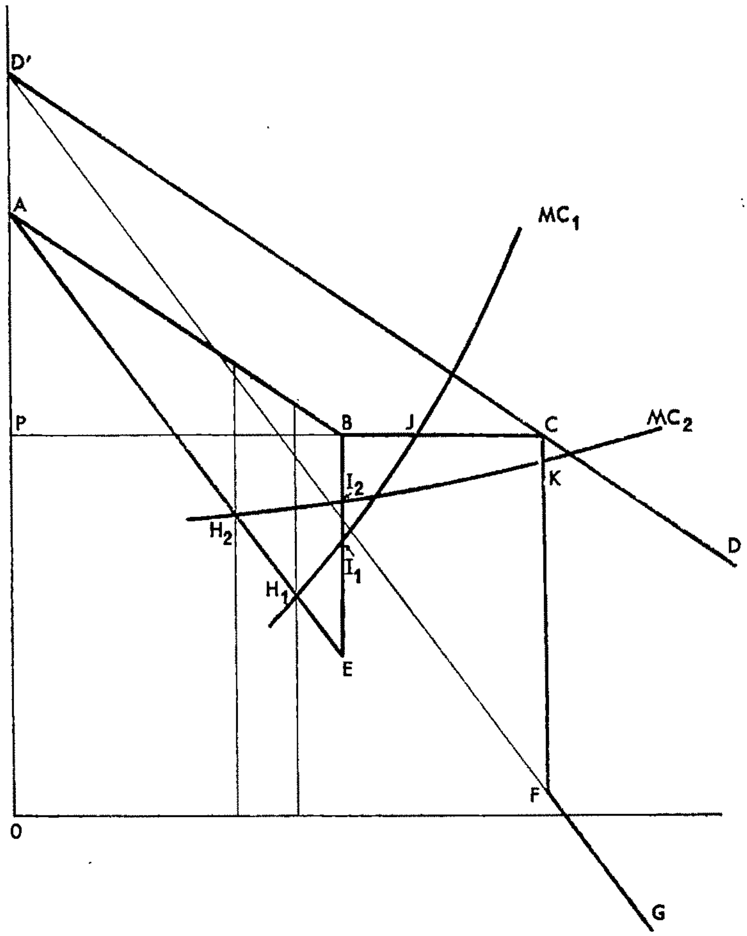


FIGURE 1

$Q<Q_1$, the quota is effective. At $Q=Q_1$, there are discontinuous changes, as the monopolist switches his maximizing policy and moves from H_1 to J (Figure 1), or from H_2 to K. As long as the quota was small enough, the domestic producer could afford to let in those limited imports and extracted the monopoly price from the residual market. But once the quota exceeds Q_1, it is more profitable for him to limit imports by cutting his price to the level of the import price. Imports fall to JC (Figure 1 with MC_1) or even to zero (with MC_2). Thus, with $Q>Q_1$ the tariff becomes effective, and further increases of the quota have no effect.

Let us look at the situation again in terms of imposing an effective quota (reducing Q). The quota frees the monopolist from the unlimited competion of imports and enables him to monopolize the residual market. It should be noted that if (as in Figure 1) MC cuts $D'G$ below the import price, *any* effective quota will reduce domestic output below the no-quota level! This case is depicted in Figure 2 (c) by the continuous line; the alternative case is shown by the broken line.

In Figure 3, the dependent variables are plotted as a function of the tariff, T. Here there are four different regions. For a low tariff, $0\geq T>T_1$, it is the quota which is effective. At $T=T_1$, there again occurs the switch from H_1 to J (or from H_2 to K). At

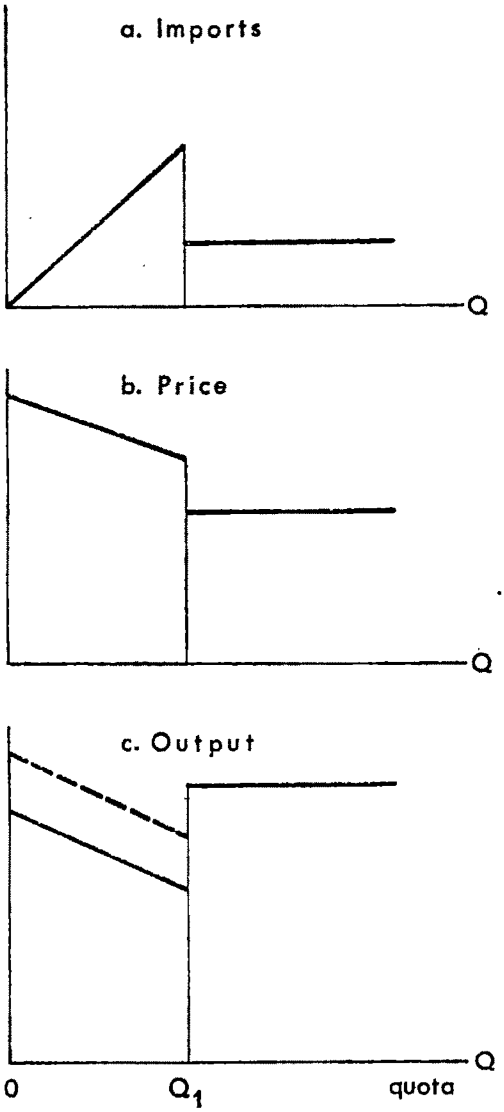


FIGURE 2

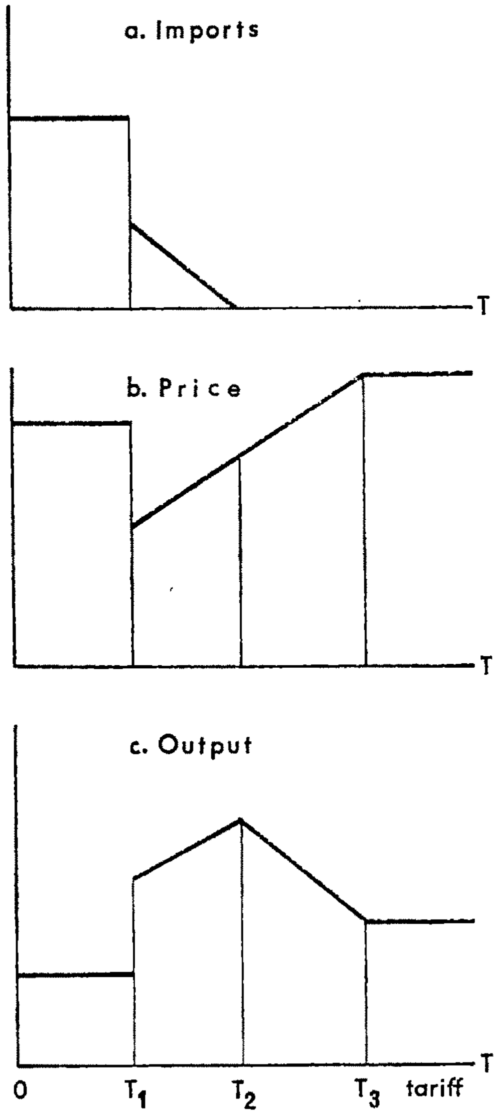


FIGURE 3

this stage, the import price has risen enough so that it pays the monopolist to cut his price to the import price level, thereby expanding his output and competing with imports; the latter are reduced (with MC_1) or even eliminated (with MC_2). With T increasing further, the point of intersection J gradually moves toward C, finally coinciding with it when $T=T_2$. With a flat enough cost curve, (MC_2), this second region would be absent. In the third region, MC intersects MR on the segment CF. Actual imports are zero, but their potential competition limits the market price to the (gradually rising) import price. This region exists independently of the imposition of a quota; here, the monopolist fixes his price so as to keep out competing imports, and any increase in the tariff enables him to raise his price accordingly, with a corresponding re-

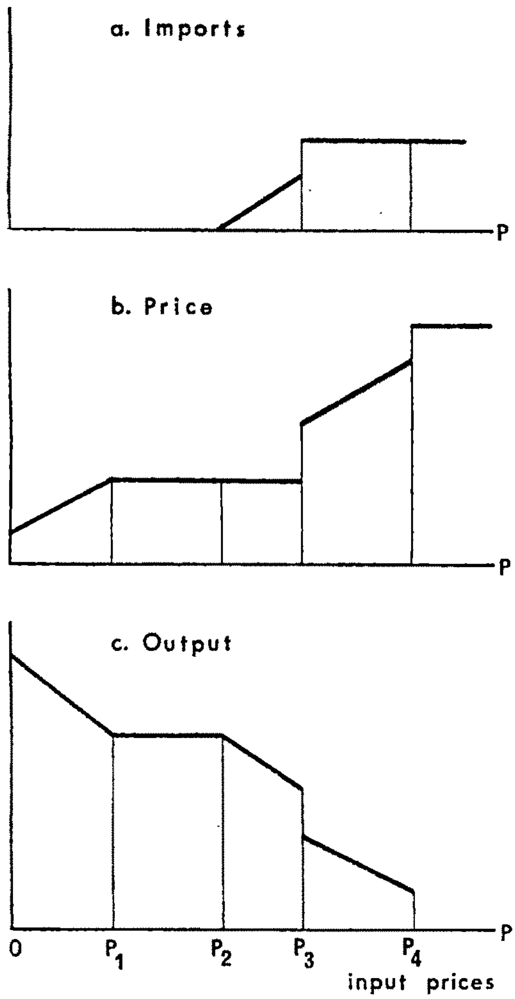


FIGURE 4

duction of his output. Finally, with $T > T_3$, imports are priced out of the market and a simple monopoly situation is created.

For the sake of completeness, Figure 4 shows the effect of changes in domestic production costs, in the form of proportional increases of input prices (P). The first region corresponds to MC intersecting MR on FG, the second to an intersection on CF, the third (which may be absent) to one on BC. At $P = P_3$, there occurs the switch from J to H_1. In the fourth region, H_1 moves upwards on AE. Finally, when $P = P_4$, domestic production is priced out of the market.

References

1. J. Bhagwati, "On the Equivalence of Tariffs and Quotas," in R. E. Baldwin and others, editors, *Trade, Growth and the Balance of Payments—Essays in Honor of Gottfried Haberler*, Chicago 1965.
2. ———, "More on the Equivalence of Tariffs and Quotas," *Amer. Econ. Rev.*, Mar. 1968, *58*, 142–46.
3. H. Shibata, "A Note on the Equivalence of Tariffs and Quotas," *Amer. Econ. Rev.*, Mar. 1968, *58*, 137–42.
4. G. Yadav, "A Note on the Behaviour of Tariffs and Quotas," *Can. J. Econ.*, Feb. 1968, *1*, 100–10.

Decentralization Under Workers' Management: A Theoretical Appraisal

By JAROSLAV VANEK*

The purpose of my paper is to carry out a theoretical appraisal of a labor-managed economy of the Yugoslav type. In my opinion, it is this type of economy that represents, by and large, the true aspirations of reformers in eastern Europe.

There is full justification for such an endeavor. On the one hand, the empirical studies which we have on the subject are not fully conclusive; on the other, even if quantitative results fulfilled some accepted criteria of statistical significance, we still might want to verify them through a theoretical evaluation. Moreover, it ought not to be forgotten that the Yugoslav experiment is unique and comparatively young. To compare its real performance with that of systems which have been tried in a large number of instances over long periods, is not quite fair. At least as a supplement to the empirical studies, it is thus necessary to compare the corresponding theoretical models.

Besides the direct and practical objective of my paper, related to the economies of eastern Europe, there is also the more academic question of the theoretical literature on the subject at hand. The latter, although of excellent quality, is extremely limited. In fact, there are only two important articles on what I would call economics of labor participation—those by Benjamin Ward [7] and Evsey Domar [3]. These authors do not make anything near a full evaluation of an entire economic system. And yet, the findings emerging from such a limited coverage often are taken as characteristic of the efficiency of the system. For example, Ward's perverse supply elasticity in the short run is taken by many as the proof of the absurdity of labor management. Perhaps the best and most authoritative illustration of the overall pessimism regarding the labor-managed economy is contained in Abram Bergson's more recent evaluation of market socialism [2].

At this stage of the argument, I do not want to dispute any specific points regarding the labor-managed economy; that will be done explicitly or by implication later in our discussion. I only want to contend that to appraise an economic system it is necessary to consider all of its major aspects. And this is what I propose to do.

The task would be an impossible one if each point were to be fully explained or proven. Fortunately, I have produced a more complete analysis elsewhere [5], [6], and thus I can restrict myself here to the presentation of the main conclusions, supplementing these with brief indications of the underlying reasoning. I will further limit my analysis by concentrating on questions of global economic efficiency, leaving out, as much as this is possible, the mechanics of the system.

I will first consider what may be referred to as a dehumanized model where, as in conventional capitalist theory, labor is considered merely a factor of production, of constant quality, exogenous to the system. This is done in Sections I and II. In Section III, by contrast, I discuss some of the most important special dimensions which emerge from the participatory nature of labor management. The principal objective of the remaining section is an overall evaluation of the labor-managed system in comparison with other major world systems.

I. *The Pure Model*

In what I call the pure model, I make basically the same assumptions as those underlying the contributions of Ward and Domar noted above. They can be summed up as implying a perfect, competitive and smooth neoclassical world in which the moving force, contrary to the capitalist situation, is maxi-

* Cornell University.

mization of income per laborer. There is only one type of labor, perfectly homogeneous, and active only as a factor of production. The only characteristic emanating from labor management is income sharing and the behavioral principle of maximization just noted. It should be clear from the outset that an economy thus defined is far from the complex reality of Yugoslavia, or any other economy adopting labor management; however, in my opinion the assumptions capture the ideal form of the economy, and thus their implications should be studied as a first step in any comprehensive evaluation.

To deemphasize the single firm approach used by Ward and Domar, it may be desirable to reverse the process, and start with the discussion of a complete full employment general equilibrium solution of the labor-managed economy (situations involving unemployment will be taken up in the next section).

Let our starting point be the ideal conditions of perfect competition and full employment. We may make the observation that when all firms of an industry use the same technology and free entry is guaranteed, the labor-managed economy will be Pareto optimal. In other words, just like its ideal capitalistic counterpart, the labor-managed economy will be producing the maximum producible output from given resources, and the maximum social satisfaction for a prescribed distribution of income.[1] These conclusions follow from the fact that competitive labor-managed firms equalize factor marginal products to factor returns for all factors including labor, from competition in nonlabor factor markets, free entry of firms and identical technologies.

By contrast, as has been pointed out by Ward [7], if technologies of different firms within an industry are different, the optimum solution will not be reached by the labor-managed economy because the equilibrium behavior will not lead to equalization of marginal value products of labor among firms. In this context the capitalist alternative appears as superior, on the assumptions made. Of course, if imperfections in the labor market are permitted, the differences in marginal productivities of labor among firms and industries may be just as important. In any case, the comparative inefficiency of the labor-managed case here discussed should not be too pronounced. In fact, the labor-managed firms are all bound to operate at a point of maximum factor productivities (i.e., where their technologies are linear-homogeneous), the comparative shortcoming being merely imputable to the fact that the more efficient ones do not produce enough. Moreover, as we argue later in Section III, labor-managed economies will normally be in a much better position to proliferate the best technology throughout the industry. Finally, even if in the case studied the labor-managed economy will produce less than the capitalist with identical factor endowments, it can be shown that it will generate a stronger demand for, and correspondingly higher returns to, capital. Thus accumulation and growth may proceed at a higher rate, leading to more output and more consumption over time.

While recognizing the abstraction and lack of realism of all of the pure theory presented in this section, it may be said that the perfectly competitive model just discussed is the least realistic. Indeed, the requirements of product homogeneity and a very large number of sellers (producers) are satisfied only in a few industries. Much more frequent is some degree of monopoly power coupled with product differentiation and, in many instances, active sales promotion. It is in this context that, I believe, Ward's earlier conclusion of greater restrictiveness of labor-managed monopolies, while formally correct, is misleading when we want to evaluate the theoretical performance of the system.

It is true that if the government or some other external agent were to set the number of firms for each industry at one, such a monopoly would be, *ceteris paribus*, more restrictive and thus socially more harmful than a capitalist monopoly. But in the real world monopolistic tendencies and market power are hardly ever of this type. Rather

[1] We will speak of the distributional effects of labor management in Section III.

they derive from the fact that efficient production is consistent with only a limited number of firms given the size of the market, the desire of firms to accumulate and grow, the artificial creation of barriers to entry, etc.

If viewed in this more realistic context of the entire industry, one can make a case for labor management which is strong indeed. On grounds of several arguments, labor management can be expected not only to yield market structures more competitive than any other free economy, but also prevent a good deal of wasteful and harmful sales promotion. The first and perhaps simplest argument is that with an increasing scale of operation, the benefits from participation—incentives, identification, and involvement—will tend to diminish.

Thus it can be expected that, all other things being equal, the point of maximum efficiency, which also is the point of long-run equilibrium for a labor-managed firm, will be reached for a lower level of output than for a firm operating with a hired labor factor. And thus, figuratively speaking, there will be room for more firms in a given industry.

The next point also hinges on the most efficient scale of operation. The labor-managed firm will never grow beyond that scale, whereas a capitalist firm often will; its growth being governed, even after greatest technical efficiency is reached, by the desire for profit maximization. Two extreme situations may further clarify the essence of this argument. Consider an ideal firm operating under constant returns to scale and facing a constant price yielding positive unit profits under capitalism. The equilibrium level of operation of that firm, if capitalist, is (at least in theory) infinite, while if labor-managed, it is finite and indeterminate at any level of operation. The simple crux of this deduction is that while the first firm must grow indefinitely to maximize profits, the second maximizes income per laborer at any level of operation. The other extreme situation is similar to that just described except that the firms in question face a less than infinitely elastic demand function. In that situation the capitalist monopoly, as is well-known, will find its equilibrium where the marginal (and average) cost reaches the marginal revenue; whereas the labor-managed firm must operate at zero output, or, more realistically, with only one employee; indeed, as the reader will easily verify, under the assumed conditions income per laborer will be maximized with only one member of the labor-managed firm. Clearly, under the assumed conditions it would take a very large number of firms to fill the industry (with a reasonable return to labor), but by the time such a large number would enter, the market power of each firm would become very low or disappear altogether.

Probably the most significant comparative advantage of the labor-managed oligopoly arises in connection with product differentiation and sales promotion. It can be shown that the labor-managed firm will in equilibrium produce less and engage in a less intensive advertising campaign than an equivalent capitalist firm as long as the latter makes positive profits, a condition which obtains virtually without exception. Especially if we realize that it is precisely the more extreme doses of exhortational advertising that constitute the heaviest social costs, the advantage of the labor-managed alternative is considerable. Not only can we expect a significantly less concentrated industry, but the external diseconomies, mind pollution, etc., should be significantly reduced or completely eliminated.

The last point deserving mention is based on the simple notion that there is a far greater desire within democratic structures to decentralize than in nondemocratic ones. For example, even with very small or no economies of scale, a capitalist, or for that matter a Soviet-type, firm will tend to retain its centralized organization, whereas the labor-managed one will attempt to subdivide itself into autonomous decision making and production units based on location or other functional characteristics. I am quite convinced that this tendency is present with the labor-managed alternative for other reasons than the stronger production-incentive of small groups.

Finally, before concluding this section, we

ought to reconsider Ward's negative supply elasticity, and discuss that and related issues in the context of general equilibrium efficiency. It is true that if short-run reactions of competitive firms were as posited by Ward, the resulting inefficiency could be quite considerable, and what is equally important, at least some markets could become unstable. The contention which I want to argue below is that matters are far from being so bad.

First of all, we have the two arguments made by Domar: One, with more than one product by the firm, even short-run elasticities can be positive and quite high for an individual product; two, even with only one product the supply elasticity in the short run must be positive if the firm operates with an (active) external labor supply constraint, reflected by a labor supply function of finite elasticity [3]. In addition, there is the problem raised by Joan Robinson [4] in her critique of Domar. Using my own words to push the argument a step further, how can one reasonably expect that a working collective will mutilate itself (kicking out say 1/10 of the membership), if it has already realized a significant gain from a price increase, say 10 percent of income, for the sake of gaining an extra, say, one percent? Indeed, this sounds like an extract from a book of rules of capitalist conduct. While what underlies the above rhetorical question certainly is true, I would further like to point out that the elimination of the negative elasticity argument does not even call for any higher morals. In the short-run, a single-product producer operating under perfect competition will have very little possibility of capital-labor substitution; in terms of Ward's diagrams, the relevant stretch of the marginal productivity of labor will generally be very near a vertical line. And, consequently, the normal short-run supply elasticity of a competitive labor-managed firm producing a single output can for all practical purposes be considered zero, even if group solidarity were absent.

Before leaving the short run, two observations are in order. First, let it be noted that the zero elasticity eliminates possibilities of instability, and reduces a good deal the loss of general equilibrium efficiency (note that in a world with only single-product firms a change in relative prices would now keep, in the short run, the point on the production possibility locus unchanged). Second, in the context of aggregate national income analysis, the zero short-run supply elasticity by no means is something to frown on; but this we will discuss in the next section.

Finally, a few words on long-run adjustments within the general equilibrium framework. Again, the introduction of an additional piece of information from the real world helps a good deal. It is reasonable to expect that with most firms the point of constant returns to scale (the point corresponding to the minimum of the long-run average cost curve) in reality is not a point, but a whole range (of "efficient outputs")—a range the end of which is hard to establish empirically. But once in such a range, as we have noted already, the long-run equilibrium of the labor-managed firms becomes indeterminate. The firms now can arbitrarily determine their scale of operation, and the most logical is that given by the size of the working collective, more or less exogenously determined. But as the reader will find easy to verify, under these conditions the long-run supply elasticities will be positive and can be quite high even with single-product firms, and the general equilibrium adjustment to changing demand conditions will be quite efficient indeed. Of course, the most efficient and full (Pareto) optimal adjustment will occur once labor incomes are equalized through entry and exit—or, with firms operating under constant returns to scale, through expansion or contraction of existing firms.

II. *Macroeconomics*

We now turn to the determination of aggregate variables such as income, employment, the price level, etc. I feel that it will be most expedient first to summarize the principal results that can be obtained for the labor-managed economy and then, to the extent that space permits, to elaborate on some of them.

The first observation is that unemployment is conceivable in a labor-managed economy, but if it occurs it will be of an entirely different nature—because of different causes and different duration—than unemployment of the Keynesian type in a capitalist economy. A second important point is that variations in effective demand can be expected in the labor-managed economy to lead primarily to variations in prices and not much or not at all to adjustments in income and employment. A third and related point is that on the whole there will be little natural cause for secular inflation, although variations in prices may be wide if markets are entirely free. Fourth, even if unemployment arises there will always be natural forces to restore full employment after a while, provided that capital markets are competitive. Fifth, the macroeconomic general equilibrium determining simultaneously the aggregate variables must be stable provided that real cost of capital varies with the general price level, and almost certainly it will be stable even if the money cost of capital is constant.

The key to points one and two is the low or zero short-run supply elasticity of firms. If for one reason or another the existing number of firms at given product and factor prices is incapable of employing (in equilibrium) all the labor force, then unemployment will prevail, and the supply elasticity being low, short-run changes in demand will be translated into price level variations. Note that this has nothing to do with wage-rigidity. Such rigidity in fact is highly unlikely if not impossible in the labor-managed economy because of the residual nature of labor income and the direct managerial ability of labor to make choices between price reductions on the one hand and idleness and sales reductions on the other. Thus, contrary to a capitalistic situation, a symmetrical flexibility of prices, up and down, and no systematic inflationary tendency should be expected.

Point four is a logical consequence of the fact that if there is unemployment, the unemployed will be in a considerably favored position in competing with existing firms for capital funds for development of new projects. Indeed, with very low unemployment incomes those currently unemployed will be able to offer considerably better terms to the lender. And thus any significant degree of unemployment carries with itself strong, even if perhaps not overly speedy, self-correcting forces. The prime vehicle of the latter is entry, and perhaps also expansion of existing firms (often into new lines of production) operating under constant returns to scale, whose scale of operation is arbitrary. It should be noted here that no parallel remedy exists for Keynesian unemployment; here either entry or expansion of existing firms implies opening of a deflationary gap and the ensuing return to a less than full employment equilibrium.

Point five asserting stability of the macroeconomic system involves some slightly involved mathematics and consequently I omit its demonstration. Full proof is contained, however, in my *General Theory of Labor-Managed Market Economies* [5].

We may sum up by taking the position of the policy maker in the labor-managed economy: He need not fear autonomous variations in demand as a cause of cyclical unemployment. Rather he may look out for such variations in order to prevent fluctuations in prices. In this endeavor he will be greatly assisted by an open and competitive foreign trade sector. Full employment does not entirely disappear, however, as the policy maker's problem. But he must be concerned with it as a matter of long range strategy, primarily ensuring perfect functioning of the capital market with proper interest structures. More actively, he may want to take discretionary steps: market research and consulting, indicative planning, forecasting, etc., to enhance the speed and efficiency of new entry and expansion. In the final analysis, however, he has a considerable advantage compared to his capitalist colleague in that matters are bound to improve even if perchance he fell asleep on the job.

III. *Special Dimensions of Labor Management*

Even if the labor-managed economy can pass with flying colors the scrutiny of conventional micro- and macro-economic theory, as presented in the preceding two sections,

its greatest strength lies in what we may identify as its *special dimensions*; dimensions largely absent in other economic systems. They are all related to, contained in, or emergent from the managerial function of labor, that is, of all participants of the enterprise. The field of analysis opened by these special dimensions is so vast that it cannot be treated in a single section; consequently we will restrict ourselves here to the outline of a few of the most important arguments. In my larger study referred to earlier, I devote to the subject several chapters [5, Part III], and even there, I certainly cannot pretend that my analysis is exhaustive.

The first and simplest consideration is that, contrary to the models studied in the two preceding sections, labor is not unique and homogeneous, but rather that in every enterprise a large number of individuals of different skills and qualifications cooperate in a common endeavor. One necessary task of labor management thus is to decide on the distributive shares among the different labor categories. The specific form of the distribution schedule, indicating whether the director gets four times the pay of the janitor, etc., will be the outcome of two sets of forces: Conditions of the labor market or more precisely, of the quasi-labor market, because there is no conventional labor market in a labor-managed economy; and the collective will of the working community as it emerges from the democratic decision making process of labor management.

Noting that the labor-managed firm in equilibrium (i.e. when maximizing income for each laborer) will equalize the income of each labor category to the corresponding marginal value product, the first set of forces will guarantee that the allocation of labor throughout the economy will be at least approximately consistent with maximum aggregate output.[2] The second set of forces, on the other hand, will guarantee a reasonable distribution of income, consistent with the generally accepted notion of justice for a particular segment of society, a particular community, or a particular period of time. This "dual" optimization of social welfare, striking a balance between the mechanistic rules of efficient resource allocation and a collective expression of distributional justice, seems to be a solution superior to that offered by either set of criteria in isolation. The cardinally important thing is that the "dual" system guarantees that major mistakes will not be committed in either the allocational or the distributional sphere.

The labor force in a labor-managed economy is not only diverse but also of variable quality. And this brings us to the second important *special dimension* of labor management. Without any doubt, labor-management is among all the existing forms of enterprise organizations the optimal arrangement when it comes to the finding of the utility-maximizing effort, i.e. the proper quality, duration and intensity of work, by the working collective. Not only is there no situation of conflict between management and the workers that might hinder the finding of the optimum, but the process of self-management itself can be viewed as a highly efficient device for communication, collusion control and enforcement among the participants.

This should be contrasted with a situation of most other enterprises where the worker normally will be furnishing a minimum effort consistent with the retention of his job. It is true that expectation of promotion or a raise may stimulate effort over the acceptable minimum, but recall that this factor should be equally operative in all firms, and thus it does not establish a comparative disadvantage of the labor-managed firm.

Our second special dimension may conveniently be summarized in terms of a diagram. In Figure 1 we measure effort (E) to the left of O along the horizontal axis, whereas on the vertical axis we find a (average) laborer's income Y. A typical transformation function for the labor-managed alternative between effort and income is indicated by the contour *ab* (note that with some fixed costs it takes some effort to generate

[2] In pure theory a frictionless system should generate perfect income equalization through entry and/or exit of firms. However, considering that in the real world there is never enough time to bring such an equalization process to completion, given imperfect information, attachments and loyalties of individuals to the working collective, etc., the competitive forces can be expected to produce only approximate results.

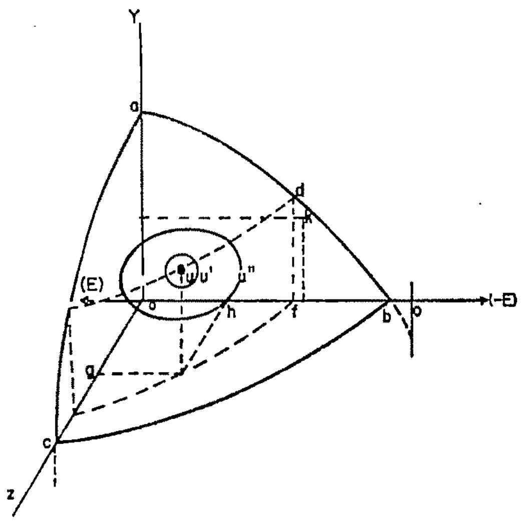


FIGURE I

zero income). By contrast, a typical transformation function in capitalist enterprise is indicated by the rectangular broken contour defined by point *k*. Its comparative deficiency is of two kinds: 1) Point *k* is below *ab*, on account of some residual profit normally not distributed among the members, and 2) and much more important, *ab* is continuous. On account of both, (as the reader may verify by drawing a set of convex indifference lines) losses in individual and collective utility, that is, in real income, will occur in one type of undertaking as compared with the other.

Third, as shown by the diagram, there is another special dimension, *Z*—in fact, it should be understood as a summary index standing for a very large number of special dimensions—also entering together with income and effort, the utility function of the members of the working collective. As indicated by the three dimensional transformation locus in Figure 1, attainment of higher levels of *Z* involves an opportunity cost in terms of income and/or effort. Nevertheless, this new dimension further augments the number of choices of the working community and thus unambiguously (as shown by the equal–utility contours u, u', u'' and the maximum-utility point *e*) its real income. It is the comparison between point *e* and point *k* that should, in the context of our present analysis, be included in any comparison between a labor-managed and a capitalist, or for that matter, Soviet-type firm.

It would hardly be possible to be exhaustive in explaining all the variables for which *Z* actually stands. However, some illustrations should be helpful: For example, the working collective, or a subgroup of it, may prefer a conveyor belt to a traditional way of intraplant transportation, even if this should involve a reduction in money income. By contrast, a capitalist would perform such a substitution only if profit could be increased. Another example is training and education of some members of the community at the expense of the enterprise. Again, the corresponding motivation and decision making is entirely different from what it would be in other firms. Still another example is collective consumption using part of the global income of the enterprise, including housing, recreation, etc. Moreover, the enterprise as a whole may undertake at its own expense social action directed toward the outside, the local community or other, and thus derive intangible real income—we may call it peace of mind—for itself.[3]

The last example brings us to a fourth category of special dimensions, related to *Z* but not exactly the same, referred to as Z'. It resembles *Z* in all respects, including that it enters the utility index, except that it does not affect money income or effort of the working community. The transformation locus with Z' substituted for *Z* would thus appear as a cylindrical surface generated by lines parallel to the Z' axis. Normally, such a surface will be truncated at some finite level of Z' and, obviously, an equilibrium will be found (with indifference loci convex everywhere) at the corresponding ridge, at a level of collective utility again unambiguously higher than at a point such as *d* with two dimensions only. To give substance to this rather technical exposition, let us give at least

[3] Some will object that all this can be done by individuals or through nonbusiness-based groupings: This is true, but it holds even for the labor-managed system and thus the potential of action by firms, the primary generators of resources, still constitutes a comparative advantage.

one example; it is a significant one. We have noted in Section I that the long-run equilibrium of a competitive labor-managed firm operating under constant returns to scale—a condition encountered in many real situations—is indeterminate with respect to scale of operation. Under such conditions, total employment and output of the firm may become our Z' variable. The working community considers it a positive utility to maximize employment in the community; the external constraint, and the ridge line of our above discussion now are given by the full employment in the community.

In the opinion of the writer, the argument just made has a good deal of relevance, and when further generalized can become even more significant especially in the context of less developed countries. The generalization is based on the recognition that in most situations—especially in manufacturing—we cannot postulate perfect competition; some market power will always be present. In such situations, however, all it takes is to reclassify Z', a zero opportunity cost variable into Z, a positive opportunity cost variable. But the absorption of unemployment in the community remains a positive good from the point of view of the working collective. Thus a solution involving higher employment than that dictated by strict income-per-laborer maximization still may be expected.

IV. *Concluding Observations*

Before attempting a summary evaluation of the labor-managed economy, two more matters ought to be taken up which normally enter the appraisal of an economic system. The first is the question of ability to mobilize national savings for accumulation. The second is the question of the capacity of generating technical progress and innovation.

Regarding the capacity to mobilize investment funds, the labor-managed economy, especially its socialist version, is in a very strong position. Noting that the labor managed firms must pay for the use of capital the corresponding value of marginal product, and taking a reasonable set of estimates of national capital stock and marginal capital productivity, the economy can generate net savings of between 20 and 30 percent of national income just by retaining capital income for reinvestment. These percentages, already among the highest currently encountered in the real world, can further be augmented through private savings, government surpluses, or as in Yugoslavia, through an obligation of firms to repay investment loans from current income.

With respect to the second question, the case for the labor managed economy is somewhat less clear cut; but in the final analysis, it would be difficult to conclude that the system in question is at a comparative disadvantage. When it comes to "small" inventive and innovative activity emerging as a side-effect of the productive process itself, the incentives, unity of purpose and ease of communication offered by labor management (and by income sharing, which is an integral part of labor-management) are conditions unequalled by any other system. On the other side of the spectrum, with respect to basic research and major scientific developments, the firms only rarely play a decisive role; thus in the present context the difference between alternative systems cannot be very important.

It is the middle of the spectrum, for "medium size" incentive and innovative activity, that would seem the most important. Here the modern capitalist corporation almost certainly has a comparative advantage with its own low-cost—and often quite abundant—funds that it can employ in research and development. On the other hand, a socialist economy, labor-managed or other, is likely to have a distinct advantage in proliferating innovations throughout the whole economy.

Let us now try to make a summary evaluation of the labor-managed system. I realize that such an endeavor must always contain a certain element of subjective judgment. This is so because summary evaluation really means constructing an aggregate ordinal index based on a combination of several evaluations, not all of the same sign; clearly, some weights must be employed and these, to a degree at least, will be subjective.

With this in mind, and basing myself on

ten years of intensive study more than on the present exposition, I cannot forego a set of strongly favorable conclusions; I cannot avoid them, even though I realize that I am contradicting the majority of our profession who have thought about the problem, and even though I may risk earning the displeasure of many.

In brief, the labor-managed system appears to me to be superior by far, judged on strictly economic criteria, to any other economic system in existence. In the sphere of allocation efficiency (concerning how well it utilizes its resources in national production), it is at worst equal to the western-type capitalist system in the context of a full-employment model (as discussed in Section I), while it is definitely superior in the context of the macro-economic model (as discussed in Section II) and in the context of its special dimensions (as discussed in Section III). On the side of the system's capability to grind out an efficient pattern of income distribution, there are also strong reasons to believe that a socialist labor-managed economy will do a better job than other market systems. This conclusion is based not only on the argument of collective decision making within the firm (as argued in Section III), but also on the fact that in the socialist economy the income share of capital, whether reinvested or not, will accrue to the society as a whole and not to a select group of individuals.

Compared to the Soviet-type command economy, the question of distributional efficiency really is empty of meaning because income distribution is decided by decree and not by some mechanism inherent in the economic system. As for allocational efficiency, far more important for our comparison because inherent in the system, we only have to recognize with Bergson [1] that the Soviet-type command economy is less efficient than market capitalism, and recall the above evaluation of the two market alternatives. In fact, even the weaker postulate of approximate equality between the two major world systems would suffice here to establish a preference for the labor-managed alternative over that of a command economy.

Two remarks less strictly economic, and stemming much more from intuition than from careful analysis, may be in order before closing this discussion: First, it seems to me that the comparative advantage of labor-managed systems becomes even stronger once we leave the strictly economic frame of reference and replace it by one that takes account of broader human values. Second, taking a very long view of world events, it seems to me that if there is a meeting ground for the presently conflicting major world systems and ideologies, it is one not too far from the system discussed here.

References

1. A. Bergson, *The Economics of Soviet Planning*, New Haven, 1964, p. 341.
2. ——— "Market Socialism Revisited," *J. Polit. Econ.*, Oct. 1967, *75*, 432–42.
3. E. Domar: "The Soviet Collective Farm as a Producer Cooperative" *Amer. Econ. Rev.*, Sept. 1966, *56*, 734–57.
4. J. Robinson: "The Soviet Cooperative Farm as a Producer Cooperative: Comment" *Amer. Econ. Rev.*, Mar. 1967, *57*, 222–23.
5. J. Vanek, *The General Theory of Labor-Managed Market Economies:* to be published.
6. ———, *The Participatory Economy: An Evolutionary Hypothesis and a Development Strategy*, to be published.
7. B. Ward, "The Firm in Illyria: Market Syndicalism" *Amer. Econ. Rev.*, Sept. 1953, *58*, 566–89.

IN MEMORIAM

ISAIAH LEO SHARFMAN

1886–1969

I. Leo Sharfman, professor emeritus of economics, University of Michigan, and former president (1945) of the American Economic Association, died on September 9, 1969 in Washington, D.C. at the age of eighty-three.

Born in the Ukraine on February 19, 1886, Professor Sharfman came to the United States in 1894. He attended the Boston Latin School and Harvard University (A.B., 1907; LL.B., 1910); while studying law, he was assistant in economics in Harvard College. In 1910 he became professor of law and political science at the Imperial Pei-Yang University in Tientsin, China. Returning to the United States in 1912, he began his long career at the University of Michigan, as lecturer in economics. He was made professor in 1914, Henry Carter Adams University Professor in 1947, and served as chairman of the department of economics from 1927 until his retirement in 1955.

Professor Sharfman participated actively in public affairs, and was a member of numerous federal boards and commissions. In 1955, he became a trustee of Brandeis University, involving himself deeply and constructively in its affairs. Brandeis awarded him an LL.D. in 1964, and held a memorial service in his honor on September 15, 1969.

Professor Sharfman's magnum opus was *The Interstate Commerce Commission* (five volumes, 1931–38), for which Harvard awarded him The James Barr Ames Prize. (Out-of-print for some time, the work has just been reissued; the first copies reached Washington the day after the author's death.) Professor Sharfman's several other books and score of journal articles also concerned primarily the public control of economic activity, and especially federal regulation of transportation. His meticulous analysis of regulatory problems reflected a profound legal knowledge as well as a keen sense of social justice. He saw the regulatory process as a means for a democratic society to bring its economic affairs under rational and orderly control.

He was a responsible leader of the University of Michigan faculty, where his wise counsel always commanded respectful consideration. Through the twenty-seven years of his department chairmanship, years of revolutionary change in economics, he respected and supported serious scholarship of whatever stamp. He administered the department's business with scrupulous care, and he brought the same loyalty, responsibility, and efficiency to his participation in the affairs of his profession, including his presidency of the association.

Dignified and rather elegant in bearing and expression, he was nevertheless a warm and delightful friend and a devoted advocate of a number of worthy causes.

NOTES

A Visiting Lecturer Program in Statistics has been organized for the seventh successive year. The program is sponsored jointly by the American Statistical Association, the Biometric Society and the Institute of Mathematical Statistics. The National Science Foundation provides partial financial support. Leading teachers and research workers in statistics, from universities, industry and government, have agreed to participate as lecturers. Lecture topics include subjects in experimental and theoretical statistics, as well as in probability theory, information theory, and stochastic models in the physical, biological, and social sciences.

The purpose of the program is to provide information to students and college faculty about the nature and scope of modern statistics, and to provide advice about careers, graduate study, and college curricula in statistics. Inquiries should be addressed to: Visiting Lecturer Program in Statistics, Department of Statistics, Stanford University, Stanford, California 94305.

The Boise Cascade Corporation Foundation, Inc. has awarded a $5000 grant to Syracuse University's College of Business Administration for graduate scholarships in urban studies. The grant will be used to provide support for minority group students entering the Syracuse University graduate program in real estate and urban development.

Controlling Monetary Aggregates, the proceedings of a conference sponsored by the Federal Reserve Bank of Boston in June 1969, will be made available after October 1. Principal papers were contributed by James S. Duesenberry, Warren L. Smith, Alan R. Holmes, Sherman J. Maisel, John H. Kareken, and James L. Pierce.

Annual Convention of New York State Economics Association (*NYSEA*) to be held on April 17/18 at State University of New York, College at Brockport. 1969/70 President of *NYSEA*: Dr. Rawle Farley, chairman, economics department, State University of New York, College at Brockport; 1969/70 Secretary-Treasurer, *NYSEA*: Dr. David Martin, chairman, economics department, State University of New York, College at Geneseo.

1970 *APHA* Meetings, Houston, Texas: Authors of papers for possible inclusion in the Medical Care Section Program of the American Public Health Association to be held 26–30 October 1970 in Houston may obtain abstract forms by contacting Dr. Rosemary A. Stevens, Program Committee, Yale School of Medicine, 60 College Street, New Haven, Connecticut 06510.

Two types of papers will be considered: Research reports and descriptions of programs or demonstrations. Deadline for submission is 15 April 1970. Selected authors will be notified in early June.

The Economics Institute is sponsored by the American Economic Association for foreign students of economics and agricultural economics beginning graduate work in the United States.

1970 session to be held from June 17–August 19 at the University of Colorado. Information and application forms may be obtained from the Director of the Economics Institute, University of Colorado, Boulder, Colorado 80302.

Deaths

Theodore Boggs, professor emeritus, University of British Columbia, June 1969.

Robert W. Burgess, May 27, 1969.

E. Vernon Carbonara, professor, New York University, Rockville Center, June 13, 1968.

Milton Fried, director of research, Amalgamated Clothing Workers of America, April 4, 1969.

J. Adair Hodges, professor emeritus, Kansas State University, June 2, 1969.

Edwin C. Johnson, Alexandria, Virginia, April 18, 1969.

Chester M. Kearney, associate professor of economics, University of Nevada, Reno, March 15, 1969.

Joy Leonard, University of Southern California, Los Angeles.

O. G. Lloyd, Cincinnati, Ohio, July 1969.

Thomas J. McCleneghan, February 1968.

John G. McGrew, June 1968.

Joseph D. Mooney, associate professor, University of Michigan, March 22, 1969.

Owen B. Murphy, Albany Junior College, September 15, 1969.

Robert M. Robinson, May 7, 1968.

Warren C. Scoville, professor of economics, University of California, Los Angeles, July 1, 1969.

Allen Solganick, assistant professor, department of economics, San Jose State College.

Samuel S. Stratton, March 1, 1969.

Frank Tannenbaum, professor emeritus, Columbia University, June 1, 1969.

John N. Webb, professor of economics and director, Institute of Behavioral Research, Florida Atlantic University, August 4, 1969.

James E. Young, professor, director of Bureau of Economic and Business Research, Kent State University, August 9, 1969.

Retirements

Lewis W. Adams, dean, School of Commerce and Administration, Washington and Lee University.

Cecil Carpenter, professor of economics, University of Kentucky, July 1969.

Denzel C. Cline, professor, Michigan State University, July 1, 1969.

Walter Cole, professor of accounting, University of Arkansas.

Dudley J. Cowden, professor of economics, University of North Carolina, June 30, 1969.

L. C. Cunningham, professor emeritus, Cornell University: Fulbright lectureship, University of the West Indies, Trinidad.

Robert Drew-Bear, professor of marketing, School of Business Administration, University of Massachusetts.

Harry Henig, professor of economics, Illinois Institute of Technology.

Manley H. Jones, professor of business management, Illinois Institute of Technology.

John B. Kohlmeyer, associate professor of agricultural economics, Purdue University, July 1969.

Harry W. Leonard, associate professor of agricultural economics, Purdue University, July 1969.

S. Sterling McMillan, professor emeritus, Case Western Reserve University: research associate, Research Program in Industrial Economics.

John Manning, professor of management, University of Louisville.

James E. Maxwell, professor, department of economics, University of Connecticut, September 1969.

Robert M. Ryan, associate professor of economics, Wichita State University, summer 1969.

Frederick W. Stamm, dean of Business School, University of Louisville.

Visiting Foreign Scholars

Theo G. M. Beekhuizen, Dutch exchange professor: College of Business, University of Oregon.

Peter F. Bell, University of British Columbia: visiting assistant professor, Yale University Economic Growth Center and University of Bridgeport.

C. Desmond Cohen, University of Sussex: visiting assistant professor of economics, University of Connecticut.

Dimitros J. Delivanis, University of Thessaloniki: department of economics, University of Iowa.

F. R. Faizi, Forman Christian College, Lahore, Pakistan: visiting professor of economics, Davidson College.

Gordon R. Fisher, Southampton University: visiting professor of econometrics, University of Western Ontario.

Colin Forster, Australian National University: visiting associate professor of economics, York University.

Shibshankar P. Gupta, University of Leeds: visiting assistant professor of economics, University of Connecticut.

Roy Harrod: visiting professor, economics department, Wharton School of Finanace and Commerce, University of Pennsylvania, spring 1970.

Branko Horvat, Yugoslav Institute of Economic Research: visiting professor of economics, American University.

Anders Klevmarken, The Industrial Institute for Economic and Social Research, Stockholm: visiting associate professor of banking and finance, University of Georgia.

E. J. Mishan, London School of Economics: visiting lecturer, department of economics, American University.

Charles S. Soper: visiting professor, economics department, Michigan State University.

Alain G. C. Van Peterssen: visiting lecturer, department of economics, Wharton School of Finance and Commerce, University of Pennsylvania.

Peter J. D. Wiles, London School of Economics: visiting professor, Dartmouth College, spring 1970.

Promotions

Lewis M. Abernathy: associate professor of economics, North Texas State University.

Alec P. Alexander: professor of economics, University of California, Santa Barbara.

Richard D. Aplin: professor, Cornell University.

Sue Atkinson: associate professor, department of economics, University of Bridgeport.

Richard D. Auster: assistant professor of economics, City College of New York.

R. W. Baguley: assistant professor, University of Western Ontario.

Alan B. Batchelder: professor of economics, Kenyon College.

Atha Beard: assistant professor, Auburn University.

John Beare: assistant professor of economics, York University.

Richard Bee: assistant professor, department of economics, Youngstown State University.

Henry W. Behrens: assistant professor of economics, Union College, Schenectady.

Thomas B. Birnberg: assistant professor of economics, Yale University.

William C. Brainard: professor of economics, Yale University.

Earl H. Brown: professor, Cornell University.

W. Keith Bryant: professor of agricultural economics, University of Minnesota.

Stanley B. Brzyski: associate professor, Rensselaer Polytechnic Institute.

John E. Buehler: associate professor of economics, University of Arizona.

Adolf Buse: associate professor, department of economics, University of Alberta.

J. M. Bushey: assistant professor, Auburn University.

Eugene A. Calabro: professor of economics, Salem State College.

Paul T. Cantrell: assistant professor of economics, Middle Tennessee State University.

Ken Carpenter: assistant professor of economics, York University.

R. L. Carson: associate professor, department of economics and accounting, Carleton University.

George L. Casler: associate professor, Cornell University.

Kang Chao: professor of economics, University of Wisconsin-Madison.

Nai-Ruenn Chen: associate professor, Cornell University.

Robert Cherry: assistant professor of economics, Case Western Reserve University.

Martin K. Christiansen: associate professor and extension economist agricultural policy, Agricultural Economics and Extension Service, University of Minnesota.

Edward Coen: professor of economics, University of Minnesota.

Doris M. Cook: professor of accounting, University of Arkansas.

Peter Costello: associate professor, department of economics, University of Bridgeport.

T. J. Courchene: associate professor, University of Western Ontario.

Jim Cutt: associate professor of economics, York University.

James D. Dana: associate professor, Lawrence University.

Edward E. Darrow: associate professor, department of agriculture economics and rural sociology, Ohio State University.

Martin J. Davidson: associate professor of economics, North Texas State University.

Jacob De Rooy: assistant professor of economics, Lehigh University.

Richard B. Du Boff: associate professor of economics, Bryn Mawr College.

Christopher W. Dungan: associate professor of accounting, University of Louisville.

Kenneth E. Egertson: assistant professor and extension economist marketing, Agricultural Economics and Extension service, University of Minnesota.

Robert J. Ellis, Jr: associate professor of economics, Southern Illinois University.

Hugh W. Folk: professor of economics and labor and industrial relations, University of Illinois.

Roderick D. Fraser: associate professor of economics, Queen's University at Kingston.

Richard M. Friedman: associate professor, department of economics, San Fernando Valley State College.

Marsha L. Geier: assistant professor of economics, Yale University.

Ismail Ghazalah: assistant professor of economics, Ohio University.

Robert S. Goldfarb: assistant professor of economics, Yale University.

M. L. Greenhut: professor, Texas A&M University.

Kanji Haitani: associate professor of economics, Southern Illinois University.

James L. Hamilton: assistant professor, department of economics, Grinnell College.

Robert L. Heilbroner: professor, department of economics, New School for Social Research.

A. J. Hill: professor, Auburn University.

John S. Hoyt, Jr: professor and program leader for special project development and coordination, Agricultural Economics and Extension Service, University of Minnesota.

Katherine Hsiao: associate professor of economics, Indiana State University.

Ernest L. Inwood: associate professor, department of economics, San Jose State College.

William B. Johnson: associate professor, department of economics, San Jose State College.

Edward D. Kalachek: professor of economics, Washington University.

Calvin A. Kent: associate professor of economics, University of South Dakota.

C. Rafiq A. Knan: senior economist, economics and statistics division, Central Mortgage and Housing Corporation, Ottawa.

Ramon Knauerhase: associate professor of economics, University of Connecticut.

Janis Labsvirs: professor of economics, Indiana State University.

John L. Landon, assistant professor of economics, Case Western Reserve University.

Herbert S. Levine: professor, department of economics, Wharton School of Finance and Commerce, University of Pennsylvania.

T. D. Little: assistant professor, Auburn University.

William F. E. Long: associate professor of economics, George Washington University.

Chung-tai Lu: professor of economics, Hamline University.

William A. Luker: professor of economics, North Texas State University.

David L. McKee: associate professor, College of Business Administration, Kent State University.

L. Merewitz: assistant professor of economics, School of Business Administration, University of California, Berkeley.

John F. Mead: associate professor of economics, University of Louisville.

Alexander Melamid: professor of economics, University College, New York University.

James R. Melvin: associate professor, University of Western Ontario.

John E. Moore: associate professor, department of agricultural economics and rural sociology, Ohio State University.

Cynthia T. Morris: professor, department of economics, American University.

Lynn R. Muchmore: assistant professor, department of economics, Grinnell College.

Thomas Muench: associate professor of economics, University of Minnesota.

Llewellyn Mullings: associate professor, department of economics, University of Bridgeport.

Esther Niemi: associate professor, department of economics, Youngstown State University.

Vahid F. Nowshirvani: assistant professor of economics, Yale University.

Ransford W. Palmer: associate professor of economics, Catholic University of America.

L. Parai: assistant professor, University of Western Ontario.

Andrew T. Pawlak: assistant professor of economics, Salem State College.

Benedict J. Pedrotti: associate professor, department of economics, San Fernando Valley State College.

Russell A. Pendergast: associate professor, department of economics, University of Alberta.

Willis L. Peterson: associate professor of agricultural economics, University of Minnesota.

Robert Piron: associate professor of economics, Oberlin College.

Gene E. Pollock: associate professor of economics, College of Wooster.

Jonas Prager: associate professor of economics, University College, New York University.

James E. Price: associate professor, department of economics, Syracuse University.

Martin Primack: associate professor, department of economics, San Jose State College.

Sevastian J. Raciti: associate professor, department of economics, Saint Peter's College.

U. L. G. Rao: assistant professor, department of economics, Dalhousie University.

Norman Rask: associate professor, department of agricultural economics and rural sociology, Ohio State University.

Edward P. Reagen: professor of economics, Skidmore College.

Frank K. Reilly: associate professor of business administration, University of Kansas.

Margaret M. Reilly: associate professor, department of economics, Saint Peter's College.

Charles Reitemeyer: associate professor, department of economics, Saint Peter's College.

Harl E. Ryder: associate professor, department of economics, Brown University.

Vladimir Salyzyn: associate professor, department of economics, University of Alberta.

Frederic M. Scherer: professor of economics, University of Michigan.

Peter Schran: professor, department of economics, University of Illinois.

T. Y. Shen: professor of economics, University of California, Davis.

J. A. Simon: professor, department of economics and business administration, University of Illinois.

A. M. Sinclair: professor, department of economics, Dalhousie University.

Arthur Smith: assistant professor, department of economics, St. Peter's College.

Eugene Smolensky: professor of economics, University of Wisconsin-Madison.

Marvin R. Snowbarger: associate professor, department of economics, San Jose State College.

Hugo Sonnenschein: professor of economics, University of Minnesota.

Charles G. Stalon: associate professor of economics, Southern Illinois University.

Robert R. Sterling: professor of business administration, University of Kansas.

Max D. Stewart: professor, department of economics, University of Alberta.

J. Dirck Stryker: assistant professor of economics, Yale University.

Raymond G. Torto: assistant professor of economics, Salem State College.

A. Dale Tussing: associate professor, department of economics, Syracuse University.

Dan Usher: professor of economics, Queen's University at Kingston.

Arley D. Waldo: professor and extension economist public affairs, Agricultural Economics and Extension Service, University of Minnesota.

Neil Wallace: associate professor of economics, University of Minnesota.

James H. Weaver: professor, department of economics, American University.

Donald Q. Webb: associate professor, department of economics, Trinity University.

Richard D. Wolff: assistant professor of economics, City College of New York.

Carole B. Yoho: assistant professor and extension specialist public affairs, Agricultural Economics and Extension Service, University of Minnesota.

Richard Zerbe: assistant professor of economics, York University.

L. S. Zudak: assistant professor, University of Western Ontario.

Administrative Appointments

John W. Allen: acting department head of economics, Texas A&M University.

Edward C. Atwood: dean, School of Commerce and Administration, Washington and Lee University.

Roman R. Andrus: head, department of marketing, insurance and transportation, College of Business Administration, University of Oregon.

W. Scott Bauman: head, department of finance and business environment, College of Business Administration, University of Oregon.

L. D. Belzung: associate dean and professor of economics, University of Arkansas.

George A. Benz: chairman, department of economics, St. Mary's University.

Walter Blass: assistant vice president, New York Telephone Company.

John M. Bonham: head, economics department, University of Tulsa.

Howard R. Bowen, University of Iowa: chairman, department of economics, Claremont Graduate School, Claremont Colleges.

William C. Brainard: acting director of the Cowles Foundation, Yale University.

Henry W. Broude: director of academic planning, Yale University.

Elmer H. Burack: professor of management, acting chairman of business and economics department, Illinois Institute of Technology.

Kevin H. Burley: acting chairman, department of economics, University of Western Ontario.

Dale E. Clement: dean, School of Business, University of South Dakota.

Kendall P. Cochran: chairman, department of economics, North Texas State University.

John G. Cross: associate chairman, department of economics, University of Michigan.

Sidney Davidson: acting dean, Graduate School of Business, University of Chicago.

George E. Delehanty: associate dean, Graduate School, Northwestern University.

Thomas F. Dernburg: chairman, department of economics, Oberlin College.

James J. Diamond: director of graduate programs in business, De Paul University.

L. Aubrey Drewry, University of Georgia: associate dean, College of Business Administration, Western Carolina University.

Ludwig M. Eisgruber: assistant head of department, research, agricultural economics, Purdue University.

Edward Foster: director of graduate studies, University of Minnesota.

Oscar R. Goodman: chairman, department of finance, Roosevelt University.

Edward Greenberg: professor, chairman, department of economics, Washington University.

Bernard Hall: associate provost, Kent State University.

Robert W. Harrington: professor, chairman, division of business and economics, Illinois Wesleyan University.

Sandra R. Henderson: chairman, department of economics, West Virginia State College.

John F. Henry: head, department of management, Auburn University.

Edwin J. Holstein: chairman, department of economics, Rensselaer Polytechnic Institute.

O. Henry Hoversten: acting head of department, business administration and economics, Capital University.

Charles W. Hultman: chairman, department of economics, University of Kentucky.

Mark J. Kasoff, Indiana University: assistant professor, acting chairman, department of economics, Antioch College.

Benjamin J. Katz: chairman, department of economics, University College, New York University.

Irving Leveson, RAND Corporation: chief research specialist, Office of Comprehensive Planning, New York City Department of City Planning.

Wesley H. Long: assistant director, Institute for Research on Land and Water Resources, Regional Analysis Center, Pennsylvania State University.

Chris Luneski: head, department of accounting and quantitative methods, College of Business Administration, University of Oregon.

Arthur E. Mace: assistant dean, Graduate School of Management and Business, University of Oregon.

Kenneth D. McIntosh: acting director, division of resource management, agricultural education, agricultural engineering, College of Agriculture and Forestry, West Virginia University.

Graeme McKechnie: chairman, department of economics, York University.

Joseph Martellaro: associate dean, Graduate School, Northern Illinois University.

John P. Miller: director, Institute of Social Science, Yale University.

James Moniz: associate professor, chairman, departments of accounting and business administration and economics, Upsala College.

Sarah S. Montgomery, associate chairman, department of economics and sociology, Mount Holyoke College.

N. H. Morse: chairman, department of economics, Dalhousie University.

Gail E. Mullin: professor, dean, College of Business Administration, Kent State University.

Edward B. Oppermann: deputy head, department of economics, United States Air Force Academy.

Daniel Orr: chairman, department of economics, University of California, San Diego.

James Owens: assistant dean, director of the undergraduate program, School of Business Administration, American University.

Andreas Papandreou: professor of economics, director of the graduate program in economics, York University.

Merton J. Peck: chairman, department of economics, Yale University.

Robert H. Persons, Jr.: acting chairman, department of economics, University of Bridgeport.

G. L. Reuber: dean, faculty of social science, University of Western Ontario.

Leonard A. Robinson: professor and head, department of accounting and finance, Auburn University.

Anthony Scaperlanda: acting chairman, department of economics, Northern Illinois University.

Martin Segal: chairman, department of economics, Dartmouth College.

Larry Sgontz: associate professor and director of graduate program in economics, University of Iowa.

Morris Silver: chairman, department of economics, City College of New York.

Donald S. Slacum: head, department of economics, Northwestern State College.

Bulkeley Smith, Jr.: chairman, department of economics and sociology, Mount Holyoke College.

Jerald R. Smith: assistant dean, assistant professor of management, School of Business, University of Louisville.

John L. Snare: chairman, department of economics, Mercer University.

Henry Solomon: chairman, department of economics, George Washington University.

H. Ellsworth Steele: assistant dean, Auburn University.

Paul Swadener: assistant dean, Undergraduate School of Business, College of Business Administration, University of Oregon.

Philip Taylor: director, Bureau of Business and Economic Research, University of Arkansas.

William D. Thompson: head, department of economics, Central Missouri State College.

J. Van Wagstaff: acting chairman, department of economics, Wake Forest University.

Ethel L. Vatter: acting associate dean, graduate education and research, College of Human Ecology, Cornell University.

Henry A. Wadsworth: assistant head, department in charge of extension, agricultural economics, Purdue University.

T. K. Warley: chairman, department of agricultural economics, University of Guelph, Jan. 1, 1970.

Kenneth P. Wichman: acting chairman, department of economics, State University of New York, Cortland.

Charles K. Wilber, chairman, department of economics, American University.

R. J. Wonnacott: chairman department of economics, University of Western Ontario.

Dieter K. Zschock: director, Economic Research Bureau, State University of New York, Stony Brook.

Appointments

Wynn A. Abronovic: assistant professor, general business and finance, School of Business Administration, University of Massachusetts.

Gardner Ackley, U.S. Ambassador to Italy: professor of economics, University of Michigan.

Esmond Adams: assistant professor, department of economics, Missouri State College.

James Adams: associate professor, department of marketing and transportation, Auburn University.

Carolyn R. Ahern: instructor of economics, University of North Dakota.

Edward Alban: instructor, department of economics, Auburn University.

Gerald Albaum, University of Massachusetts: associate professor of marketing, College of Business, Administration, University of Oregon.

M. Mukhtar Ali: assistant professor of economics, University of Kentucky.

Victor F. Amann, College of Agriculture and Forestry, West Virginia University: senior lecturer, Makerere University, Uganda.

Troy Anders: assistant professor, department of economics, Auburn University.

Robert J. Anderson, National Air Pollution Control Administration: assistant professor of economics, Purdue University.

Vincent P. Apilado: assistant professor, School of Business Administration, Arizona State University.

D. A. L. Auld, University of Guelph: assistant professor, public finance, University of British Columbia.

Scott Avary: instructor of economics, Mercer University.

Duncan M. Bailey: instructor, department of economics, Virginia Polytechnic Institute.

Trevor Bain, University of Michigan: visiting associate professor, Queens College of the City University of New York.

Billy W. Balch: associate professor of economics, Middle Tennessee State University.

Mary K. Barber: assistant professor of marketing, School of Business Administration, University of Massachusetts.

William F. Barnes: assistant professor of economics, University of Georgia.

Howard N. Barnum: instructor of economics, Dartmouth College.

D. Marshall Barry: assistant professor of economics, division of social sciences, New College.

Robert L. Basmann, Purdue University: professor, department of economics, Texas A&M University.

Raveendra N. Batra: assistant professor of economics, Southern Illinois University.

Dayal S. Batth, Texas Southern University: associate professor of finance, Chico State College.

Ronald Bearden: instructor, department of economics, University of Houston.

Robert J. Beatty: assistant professor of economics, United States Air Force Academy.

Sara Behman, University of California, Berkeley: assistant professor, department of economics, San Jose State College.

Robert D. Beland: assistant professor of economics, United States Air Force Academy.

J. N. Benson: assistant professor, department of economics, Wellington College, University of Guelph.

Roger R. Betancourt: assistant professor of economics, University of Maryland.

Kul B. Bhatia: lecturer, University of Western Ontario.

Kelly J. Black, Planning Research Incorporated: assistant professor of quantitative analysis, Chico State College.

Harry Block: instructor of economics, Illinois Institute of Technology.

Dwight M. Blood: professor of economics, University of Wyoming.

Lawrence P. Bohl: assistant professor, agricultural economics, Purdue University.

John J. Bonsignore: assistant professor, general business and finance, School of Business Administration, University of Massachusetts.

John M. Borcich, University of Washington: assistant professor, department of economics, University of British Columbia.

M. D. Bordo: assistant professor, department of economics and accounting, Carleton University.

Melvin B. Borland, Kentucky Southern College: assistant professor of economics, University of Richmond.

Tim Bornstein: associate professor of management, School of Business Administration, University of Massachusetts.

John Bossons: research staff associate, National Bureau of Economic Research.

J. Carroll Bottum: professor of agricultural economics, Purdue University.

Colin I. Bradford, Jr., Columbia University: research economist, Overseas Development Council.

Robert L. Brittingham, Saint Louis University: instructor of economics and statistics, St. Ambrose College.

Charles R. Broshous, Jr.: instructor of economics, department of social sciences, United States Military Academy.

Vartkes L. Broussalian, Bureau of Budget: visiting professor of economics, Duke University.

Robert C. Brown: instructor, department of economics, Northwestern State College.

Barbara C. Buddington: instructor of economics, University of Connecticut.

Vivian A. Bull: instructor of economics, Drew University.

John G. Burch, Jr.: assistant professor of accounting, School of Business Administration, University of Massachusetts.

David F. Burgess: lecturer, University of Western Ontario.

Edwin T. Burton III: assistant professor, department of economics, Cornell University.

William L. Call: assistant professor, department of economics, Northwestern State College.

N. E. Cameron: assistant professor, department of economics, University of Manitoba.

Rondo Cameron, University of Wisconsin: professor, department of economics, Emory University.

John Campbell: assistant professor of economics, Missouri Valley College.

Cornelis A. Cannegieter, American University in Cairo: visiting professor of economics, Allegheny College.

Valdemar Carlson, Antioch College: professor of economics, Iowa Wesleyan University.

Robert H. Chandross: department of economics and finance, Bernard M. Baruch College.

James E. Ciecka: assistant professor, department of economics, De Paul University.

David H. Ciscel: instructor in economics, Drake University.

C. Scott Clark: lecturer, University of Western Ontario.

Joseph D. Coffey, University of California, Berkeley: programming evaluation and planning staff, U.S. Department of Agriculture.

Robert Cohen, Brooklyn College: economist, Federal Reserve Bank of New York.

Edward T. Conmy III: lecturer in investments, department of economics, University of North Dakota.

Muriel Converse, Wayne State University: assistant professor, University of Michigan.

William A. Cox: assistant professor of economics, University of Maryland.

Francis C. Cushing, Jr.: associate professor of management, University of Louisville.

J. C. Danielsen: assistant professor, department of economics and accounting, Carleton University.

Ralph C. d'Arge, University of New Mexico: assistant professor, department of economics, University of California, Riverside.

K. Hans-Joachim Daumer: associate professor of management, Chico State College.

Joe C. Davis: assistant professor, department of economics, Trinity University.

Ernest H. Day: assistant professor of economics, University of Maryland.

Edwin A. Deagle: instructor of economics, department of social sciences, United States Military Academy.

Robert J. R. de Cotret: Council of Economic Advisers, Executive Office of the President.

Judith D. Dernburg: lecturer in economics, Oberlin College.

Barend A. DeVries: adjunct professor of economics, George Washington University.

Carlos Díaz-Alejandro, University of Minnesota: professor of economics, Yale University.

William M. Dickson: associate professor, department of economics and commerce, Simon Fraser University.

Steven W. Dobson: instructor in economics, Dartmouth College.

William D. Dobson, Purdue University: assistant professor, agricultural economics, University of Wisconsin.

Donald H. Dohrmann: instructor of economics, University of Connecticut.

Norman Douglas, Temple University: assistant professor, University of Bridgeport.

Edwin G. Dolan: instructor of economics, Dartmouth College.

Paul B. Downing, assistant professor, department of economics, University of California, Riverside.

Guy Douyon: assistant professor of business management and managerial economics, School of Business Administration, American University.

Stephen P. Dresch: research staff associate, National Bureau of Economic Research.

Alan I. Duchan, University of Wisconsin: instructor, department of economics, University of Illinois.

John Duchman, Vanderbilt University: assistant professor of economics, University of Tennessee, Chattanooga.

Wallace H. Duncan: instructor in economics, department of social sciences, United States Military Academy.

William C. Dunkelberg: assistant professor of business economics, Graduate School of Business, Stanford University.

Douglas M. Dunn: Bell Telephone Laboratories, Inc., Murray Hill, N. J.

Katherine M. Dusak: instructor of finance, Graduate School of Business, University of Chicago.

William R. Eadington: assistant professor of economics, University of Nevada.

Roberto P. Echeverria: economist, International Bank for Reconstruction and Development, Washington.

Stahrl W. Edmunds: lecturer of economics, University of California, Riverside.

Robert Eisner: research staff, National Bureau of Economic Research.

C. W. Ekstrand: assistant professor, department of economics, University of Manitoba.

Paul Eldridge: assistant professor, general business and finance, School of Business Administration, University of Massachusetts.

Amin A. Elmallah: lecturer of accounting and quantitative analysis, Chico State College.

Elizabeth Erickson: instructor, department of economics, University of Akron.

Robert G. Evans: assistant professor, department of economics, University of British Columbia.

Robert E. Evenson, University of Minnesota: associate professor of economics, Yale University.

Lattee A. Fahm, University of Lagos: associate professor of economics, University of Connecticut.

Berkwood M. Farmer: instructor of economics, department of social sciences, United States Military Academy.

Clyde H. Farnsworth, Virginia Polytechnic Institute: economist, research department, Federal Reserve Bank of Richmond.

John C. H. Fei, Cornell University: professor of economics, Yale University.

Robert L. Ferrar: assistant professor of economics, Allegheny College.

John F. Ficks, Southern Illinois University: assistant professor of economics, University of Evansville.

Ronald Filante: assistant professor, Queens College of the City University of New York.

Joseph M. Finger, Duke University: associate professor, University of Georgia.

Gordon R. Fisher: visiting professor of economics, University of Western Ontario.

William Fleischman: assistant professor of economics, division of social sciences, New College.

Terry G. Foran, Pennsylvania State University: instructor, Southern Illinois University.

Robert H. Forste: staff associate, New England Center for Continuing Education, University of New Hampshire.

Charles Freedman: instructor, department of economics, University of Minnesota.

David Fretz: instructor, division of social science, Northeastern State College.

Stephan Friedberg: visiting lecturer, University of Western Ontario.

Norbert W. Frische: instructor of economics, department of social sciences, United States Military Academy.

Charles J. Gallagher: assistant professor of economics, University of Georgia.

Charles W. Gallagher: assistant professor of economics, University of North Dakota.

George Garvy: professor of finance, Graduate School of Business, Stanford University.

F. Dermot Gately: visiting lecturer of economics, University of Western Ontario.

Paul Gerhart: assistant professor of labor and industrial relations, University of Illinois.

Gerald B. Goodchild: instructor in economics, department of social sciences, United States Military Academy.

David M. Gordon: lecturer, department of economics, Yale University.

Paul W. Gordon: associate professor of marketing, School of Business Administration, University of Massachusetts.

Daniel A. Graham: assistant professor of economics, Duke University.

J. A. Gray: assistant professor, department of economics, University of Manitoba.

Jean M. Gray: assistant professor of economics, Rutgers–The State University.

Leroy J. Grossman, Center for Urban Programs: associate professor, department of economics, Saint Louis University.

Gladys Gruenberg: assistant professor, department of economics, Saint Louis University.

Fred S. Grygiel: lecturer, department of economics, De Paul University.

Joseph P. Guiltinan: assistant professor of marketing, School of Business Administration, University of Massachusetts.

Benton E. Gup, Federal Reserve Bank of Cleveland: associate professor, University of Tulsa.

Alan L. Gustman: assistant professor of economics, Dartmouth College.

Sheldon E. Haber: associate professor of economics, George Washington University.

Hamid Habibagahi, University of Pennsylvania: visiting associate professor, University of Waterloo.

Jack O. Hakkila: assistant professor, department of economics, University of Alaska.

William K. Hall: assistant professor of statistics, Graduate School of Business Administration, University of Michigan.

Eric A. Hanushek: assistant professor of economics, United States Air Force Academy.

Robert C. Haring: professor of marketing and coordinator of Graduate Studies in Business Administration and Economics, University of Arkansas.

John M. Hartwick: assistant professor of economics, Queen's University at Kingston.

Curtis Harvey: associate professor of economics, University of Kentucky.

John A. Haslem, University of Wisconsin: associate professor, College of Business and Public Administration, University of Maryland.

Andrew Jong-Chol Hau: associate professor of economics, Millersville State College.

Thomas M. Havrilesky: associate professor of economics, Duke University.

Michael A. Heilperin, University of Southern California: visiting professor, department of economics, California State College, Long Beach.

Horst Herberg: visiting assistant professor, University of Western Ontario.

David Hershfiel: assistant professor, department of economics, University of Utah.

Richard S. Higgins: assistant professor of economics, University of Georgia.

Sydney H. High: instructor, Mississippi State University.

Seiko Hija: visiting assistant professor, department of economics, Purdue University.

C. Russell Hill: assistant professor, University of Michigan.

Harriet M. Hinck: instructor of economics, Drew University.

William T. Hogan, Fordam University: visiting professor of mineral economics, Pennsylvania State University.

William Holley: instructor, department of economics, Auburn University.

Thomas A. Hoopengardner: instructor, department of economics, Marshall University.

Carl B. Housley: assistant professor of economics, Valparaiso University.

E. Philip Howrey: associate professor, department of economics, Wharton School of Finance and Commerce, University of Pennsylvania.

E. Kay Hunt: visiting professor, department of economics, University of California, Riverside.

Lionel R. Ingram, Jr.: instructor in economics, department of social sciences, United States Military Academy.

William Jaffe, Northwestern University: visiting professor, economic history, University of British Columbia.

John James: assistant professor of economics, Middle Tennessee State University.

Dean Jamison: assistant professor, department of economics, Stanford University.

Jerry J. Jasinowski: assistant professor of economics, United States Air Force Academy.

N. T. Jazairi: assistant professor, department of economics, York University.

George D. Johnson: assistant professor of quantitative analysis and management, Chico State College.

O. Maurice Joy: assistant professor of business administration, University of Kansas.

John Kagel, Purdue University: assistant professor, department of economics, Texas A&M University.

Stephan F. Kaliski, Carleton University: professor of economics, Queen's University at Kingston.

Philip Karst: assistant professor of economics, University of Kentucky.

George G. Kaufman: visiting senior lecturer, University of Southern California, spring 1970.

Suheil Kawar: assistant professor, department of economics, San Fernando Valley State College.

Alexander K. Kelly: visiting assistant professor, University of Western Ontario, Jan. 1970.

William J. Kelly: assistant professor of economics, University of Georgia.

Edward J. D. Ketchum: lecturer, The University of Western Ontario.

Duane Kexel: instructor, department of economics, Grinnell College.

M. Saleem Khan: assistant professor, department of economics, University of Alaska.

Levis A. Kochin: visiting lecturer in economics, University of Western Ontario.

Krishna L. Kool: instructor, department of economics, Marshall University.

Benedikt Korda: associate professor, department of economics, University of Alberta.

David Kreage: research staff associate, National Bureau of Economic Research.

S. Kumon: associate professor, department of economics and accounting, Carleton University.

Donald M. Lamberton: professor of economics, research assistant, industrial economics, Case Western Reserve University.

James C. Lampe: assistant professor of accounting, Graduate School of Business Administration, Michigan State University, Jan. 1970.

John S. Lane: visiting assistant professor of economics, Queen's University at Kingston.

Kazimierz Laski: professor, department of economics, York University.

Charles R. Laudor: assistant professor, department of economics, Connecticut College.

Dennis R. LaVigne: instructor of economics, Illinois Wesleyan University.

Stephen D. Lewis: assistant professor, department of economics, University of Alberta.

John M. Leyes, Iowa State University: assistant professor of economics, University of Wyoming.

Stephen E. Lile: assistant professor of economics, University of Richmond.

Joseph A. Litterer: professor of management, School of Business Administration, University of Massachusetts.

A. J. Little, St. Patrick's College: associate professor, department of economics and accounting, Carleton University.

Stanley G. Long: assistant professor, Lawrence University.

William F. Lott: assistant professor of economics, University of Connecticut.

Milton Lower: assistant professor, department of economics, University of Houston.

Lawrence Lynch: adjunct assistant professor of economics, University of Kentucky.

Bruce R. McKellips: assistant professor, department of economics, University of Alberta.

Eugene P. McLoone: lecturer of economics, University of Maryland.

Ben R. McManus, Purdue University: assistant professor, agricultural economics, University of Tennessee.

John R. McNamara: instructor of economics, Union College.

Elizabeth C. MacRae: assistant professor of economics, University of Maryland.

Kevin L. Maher: instructor of economics, department of social sciences, United States Military Academy.

Walter R. Mahler, Jr.: assistant professor, department of economics, Grinnell College.

Gail E. Makinen: assistant professor, department of economics, Wayne State University.

Arthur Mandel, University of California, Davis: department of economics and commerce, Simon Fraser University.

Aurora S. Marcus: assistant professor of accounting and business administration, Upsala College.

Bernard J. Marks, Southern Illinois University: professor of economics, Wichita State University.

John A. Marlin: department of economics and finance, Bernard M. Baruch College.

David D. Martin: visiting professor, department of economics, Michigan State University.

John M. Martin, University of Washington: assistant professor of economics, California State College at Hayward.

Otto B. Martinson: associate professor of economics, United States Air Force Academy.

Prithvi Mathur: instructor, department of economics, University of Akron.

J. Douglas May: lecturer, University of Western Ontario.

Harry G. Meyers: instructor, Rensselaer Polytechnic Institute.

Robert T. Michael: research staff associate, National Bureau of Economic Research.

H. Laurence Miller, Jr.: research fellow, National Bureau of Economic Research.

Donald Milley: assistant professor of economics, Youngstown State University.

Franklin D. Mills, University of Chicago: acting assistant professor, department of economics, University of Washington.

Peter Miovic: assistant professor, department of economics, Cornell University.

Leonard J. Mirman: assistant professor, department of economics, Cornell University.

William R. Moffat: assistant professor of business economics, Graduate School of Business, Stanford University.

Donald R. Monath: associate professor of finance, University of Louisville.

Tait Montague: professor, department of economics, York University.

Herbert E. Moody: instructor of economics, department of social sciences, United States Military Academy.

John C. Moorhouse: assistant professor of economics, Wake Forest University.

Oskar Morgenstern: professor, department of economics, Texas A&M University.

Richard D. Morgenstern, University of Michigan: instructor of economics, American University.

Ula K. Motekat: assistant professor of accounting, School of Business Administration, University of Massachusetts.

Timothy D. Mount, University of California, Berkeley: assistant professor, department of agricultural economics, Cornell University.

Ronald E. Müller: instructor, department of economics, American University.

Wladimir Naleszkiewicz: department of economics, Cleveland State University.

Edward J. Nell, University of East Anglia, Norwich: professor of economics, New School for Social Research.

Karl H. Niebyl, Temple University: visiting professor of economics, University of Connecticut.

Lowell E. Nordquist, Western Michigan University: instructor of economics, Clemson University.

Sumiye Okubo: economist, research department, Federal Reserve Bank of Richmond.

John P. Olienyk: instructor of economics, University of South Dakota.

Alan L. Olmstead, University of Wisconsin: acting assistant professor, University of California, Davis.

Josephine Olson: department of economics and finance, Bernard M. Baruch College.

Guy Orcutt: visiting professor of economics, Yale University.

Yoshihiko Otani: assistant professor, department of economics, University of Minnesota.

Anthony Pappas, Yale University: instructor, department of economics, University of Illinois.

Chandra P. Pathak: assistant professor, department of economics, Marshall University.

Ivor Pearce: visiting professor, department of economics, Wharton School of Finance and Commerce, University of Pennsylvania.

Fred Peery, University of Texas: associate professor of marketing, University of Tennessee, Chattanooga.

Adamantios A. Pepelasis: professor, department of economics, University of Kentucky.

Joseph J. Persky: lecturer, department of economics, Yale University.

Anthony C. Petto: assistant professor, department of economics, De Paul University.

P. Phillips: assistant professor, department of economics, University of Manitoba.

Hal Pickle: professor, department of management, Auburn University.

DeWayne J. Piehl: assistant professor of business administration, School of Business Administration, University of Washington.

Thomas Pinfold: assistant professor, department of economics, Dalhousie University.

Richard H. M. Plain: assistant professor, department of economics, University of Alberta.

Charles G. Plourde: lecturer, University of Western Ontario.

Clayne Pope: assistant professor, University of Michigan.

Robert Posner: assistant professor of economics, Oberlin College.

Timothy W. Pyron: assistant professor, department of economics, Trinity University.

Assaf Razin: assistant professor of economics, University of Minnesota.

Mark J. Riedy: assistant professor, School of Business Administration, University of Colorado.

A. R. Ritter: assistant professor, department of economics and accounting, School of International Affairs, Carleton University.

Howard H. Roberts: instructor in economics, department of social sciences, United States Military Academy.

Paul O. Roberts: research staff associate, National Bureau of Economic Research.

Raymond B. Roberts: assistant professor, department of economics, Iowa State University.

James W. Robertson, School of Business, Seattle University: professor of accounting and coordinator of graduate studies in business, Chico State College.

Frank Roseman, University of Western Ontario: department of economics, University of Alberta.

John C. R. Rowley: assistant professor of economics, Queen's University at Kingston.

Brian Rungeling: assistant professor, department of economics, Auburn University.

James M. Ruth: instructor in economics, department of social sciences, United States Military Academy.

Jay S. Salkin: assistant professor of economics, Duke University.

John S. Salter: assistant professor, agricultural economics and economics, Montana State University.

Larry B. Sawers, University of Michigan: assistant professor, department of economics, American University.

Dennis E. Schlais: assistant professor of finance, Chico State College.

Maria Schmundt: lecturer, University of Western Ontario.

Erich Schneider: visiting professor, department of economics, Wharton School of Finance and Commerce, University of Pennsylvania.

David L. Schulze: assistant professor, department of economics, Iowa State University.

Morton D. Schwartz: instructor, Lawrence University.

James G. Scoville, Harvard University: associate professor of economics and labor and industrial relations, University of Illinois.

Edward J. Sedlmeier: assistant professor of economics, University of Georgia.

David Segal: assistant professor of economics, Oberlin College.

Richard G. Segers: visiting associate professor of management science, Graduate School of Business, University of Chicago.

J. R. Seldon: assistant professor, department of economics, University of Manitoba.

Perry Shapiro: lecturer, department of economics, University of California, Santa Barbara.

Jerry A. Sharples: assistant professor, agricultural economics, Purdue University.

Gloria Shatto: associate professor, department of economics, University of Houston.

Dorothy Sherling: instructor, department of economics, Auburn University.

B. K. Short: assistant professor, department of economics, Wellington College, University of Guelph.

Manuel Siguenza, University of Pennsylvania: assistant professor, University of Waterloo.

George Simmons: assistant professor, University of Michigan.

Vladimir Simunek: associate professor, College of Business Administration, Kent State University.

Paul R. Siragusa: instructor, department of economics, Northwestern State College.

Burman Skrable: instructor, department of economics, State University of New York, Cortland.

Keith Sloane: visiting lecturer, department of economics, Wharton School of Finance and Commerce, University of Pennsylvania.

J. Edward Smith, Jr.: assistant professor of economics, Union College.

Leah H. Smith: instructor of economics, North Carolina State University.

Charles Snow: visiting associate professor, department of management, Auburn University.

Donald C. Snyder: instructor of economics, Illinois Wesleyan University.

Lewis C. Solomon: research fellow, National Bureau of Economic Research.

Allen R. Soltow, Iowa State University: assistant professor of economics, University of Tulsa.

Inbum Song: assistant professor of economics, University of Louisville.

Robert C. Springer: lecturer of economics, Rutgers–The State University.

M. Lynn Spruill: assistant professor, School of Business Administration, University of Kentucky.

George J. Staubus: visiting professor of business administration, University of Kansas.

William Stober: associate professor of economics, University of Kentucky.

Anthony Stocks, West Virginia University: professor of economics, Youngstown State University.

Vladimir Stoikov, University of Illinois: professor, New York State School of Industrial and Labor Relations, Cornell University.

Richard L. Stroup: assistant professor, agricultural economics and economics, Montana State University.

Richard F. Sullivan: assistant professor, department of economics and accounting, Carleton University.

Gerald M. Surett, Bennington College: instructor, department of economics, State University of New York, Cortland.

Roger M. Swagler: instructor of economics, Drake University.

Craig Swan: instructor, department of economics, University of Minnesota.

Lawrence B. Swartz: assistant professor of economics, University of North Dakota.

Koji Taira, Stanford University: professor, Institute of Labor and Industrial Relations, University of Illinois.

Paul J. Taubman: research staff associate, National Bureau of Economic Research.

Harshadray H. Thaker, Appalachian State University: instructor in economics, Clemson University.

R. Thakkar: assistant professor, department of economics, York University.

Kenneth H. Thomas: visiting assistant professor, agricultural economics, Purdue University.

Emanuel Tobier: associate professor, Queens College of the City University of New York.

Yasushi Toda: assistant professor, University of Western Ontario.

Jay L. Tontz, United States Air Force Academy: assistant professor of economics at California State College, Hayward.

Richard J. Trethewey, University of Washington: instructor, Kenyon College.

S. C. Tsiang, University of Rochester: professor, department of economics, Cornell University.

Norman B. Ture: adjunct professor of economics, George Washington University.

Robert G. Uhler: assistant professor of economics, department of social sciences, United States Military Academy.

John J. Valentini: lecturer, department of economics, De Paul University.

Charles D. Van Eaton: assistant professor, department of economics, Northwestern State College.

R. D. Vanderberg: assistant professor, department of economics and accounting, Carleton University.

Gerald Varble: instructor, division of social science, Northeastern State College.

David A. Vardy: assistant professor of economics, Queen's University at Kingston.

Claude Vaughan: instructor of economics, University of Kentucky.

Annibal Villela: visiting professor of economics, Vanderbilt University.

Donald F. Vitaliano: assistant professor, Rensselaer Polytechnic Institute.

Howard Wachtel, University of Michigan: assistant professor, department of economics, American University.

Charan D. Wadhva: assistant professor of economics, Illinois Institute of Technology.

Terence J. Wales: research staff associate, National Bureau of Economic Research.

Roger W. Walker: professor of economics, Hamline University.

Robert Watkins: instructor, department of marketing and transportation, Auburn University.

Phillip Weitzman: assistant professor of economics, University College, New York University.

Kenneth Wertz: assistant professor, University of Michigan, winter 1970.

Donald A. West, University of Wisconsin: assistant

professor, department of agricultural economics, Washington State University.

Obie G. Whichard: assistant professor of economics, University of Georgia.

David A. Whitaker, University of Florida: assistant professor of economics, University of Richmond.

Ray D. Whitman: lecturer and reserach associate, Bureau of Business and Economic Research, University of Maryland.

James L. Wiek: assistant professor of marketing, School of Business Administration, University of Massachusetts.

Ernest D. Wilcox: instructor, University College, New York University.

E. John Williams: instructor, department of economics, University of Houston.

John W. Wilson: instructor in economics, department of social sciences, United States Military Academy.

David A. Wilton: assistant professor of economics, Queen's University at Kingston.

Albert G. Wise, Memphis State University: instructor of economics, Clemson University.

Andrew J. Winnick: acting assistant professor of economics, University of California, Santa Barbara, Jan. 1970.

Alexander C. Wiseman: assistant professor, Utah State University.

William W. Wood: lecturer in economics, University of California, Riverside.

Allen D. Woodland, University of New England, New South Wales: assistant professor, department of economics, University of British Columbia.

Gavin P. Wright: lecturer, department of economics, Yale University.

Benno Wymar: assistant professor of economics, University of South Dakota.

Thomas B. Yandle, Jr., Georgia State University: assistant professor of economics, Clemson University.

Tae-Hee Yoon, Lakehead University, Ontario: senior economics, Planning and Evaluation Branch, Canada Department of Manpower and Immigration, Ottawa.

George E. Young: instructor of economics, department of social sciences, United States Military Academy.

John H. Young, University of British Columbia: chairman of the Prices and Incomes Commission, Government of Canada.

Muhammad Yunus: assistant professor of economics, Middle Tennessee State University.

Ezzatollah Zanganeh: assistant professor of economics, West Virginia State College Institute.

Joseph A. Ziegler, University of Notre Dame: assistant professor of economics, Clemson University.

Martin E. Zweing: department of economics and finance, Bernard M. Baruch College.

Leaves for Special Appointments

Martin E. Abel, University of Minnesota: program advisor in agricultural economics, Ford Foundation, New Delhi.

D. J. Allee, Cornell University: Systems Analysis Group of the Special Assistant for Civil Functions, Office of the Secretary of the Army.

Julian H. Atkinson, Purdue University: chief of party, Purdue-Brazil Project, Vicosa, Brazil.

Richard E. Attiyeh, University of California, San Diego: University of Edinburgh.

Emerson M. Babb, Purdue University: visiting professor, University of California, Berkeley.

David Baerncopf, College of Business Administration, University of Oregon: Bank of America, San Francisco.

Francis R. Bethlen, College of Arts and Sciences, State University of New York, Plattsburgh: visiting professor, lecturer of economics, University Rosario, Argentina.

T. M. Brown, University of Western Ontario: department of economics, University of Ghana.

Edward E. Carson, Purdue University: visiting professor, University of California, Berkeley.

Kee Il Choi, College of William and Mary: visiting associate professor of economics, University of Miami.

M. O. Clement, Dartmouth College: Fulbright lecturer, Robert College, Istanbul.

John C. Dawson: Rockefeller Foundation teaching and research program, Kampala University, Makere.

Floyd Durham, Texas Christian University: Danforth fellowship, Yale University.

Martin T. Farris, Arizona State University: visiting professor of economics, University of Hawaii.

George R. Feiwel, University of Tennessee: visitor, faculty of economics, University of Cambridge.

Douglas Fisher, Queens College of the City University of New York: temporary senior lecturer, University of Essex.

Irving Gershenberg, University of Connecticut: visiting assistant professor of economics, Makerere University College.

Alan L. Ginsburg, The City College of New York: fellowship, Division of Health, Education and Welfare, The Brookings Institution.

Bruce Glassburner, University of California, Davis: associate director, Hong Kong Study Center, University of California, 1969–70; director, 1970–71.

Alan H. Gleason, International Christian University, Japan: visiting professor, University of Pittsburgh.

Ronald B. Gold, Ohio State University: fiscal economist, Tax Analysis Office, Treasury Department.

Donald F. Gordon, Rochester University: Rockefeller Foundation teaching and research program, University of Philippines, Quezon City.

Douglas H. Graham, Vanderbilt University: University of Sao Paulo, Brazil,

Albert L. Gray, Jr., Baldwin Wallace College: visiting professor, American University, Cairo.

George Green, University of Minnesota: assistant visiting professor, Harvard University.

Mitchell Harwitz, State University of New York: Rockefeller Foundation teaching and research program, University College, Nairobi.

Mark A. Haskell, New York University: teaching and research program, Makerere University College, Kampala.

John Hause, University of Minnesota: National Bureau of Economic Research.

Robert H. Haveman, professor of economics, Grinnell College: Joint Economic Committee of Congress.

James Heilbrun, Columbia University: Ford faculty fellow.

George H. Hildebrand, Cornell University: Deputy Undersecretary for International Affairs, Department of Labor.

Clifford Hildreth, University of Minnesota: visiting professor, Virginia Polytechnic Institute until March 1970, visiting professor, University of Tokyo, March–August 1970.

Karel Holbik, Boston University: visiting professor, Institut d'Etudes Europeennes, Universite Libre de Bruxelles, Brussels, Belgium.

Hendrik Houthakker, Harvard University: Council of Economic Advisors, January, 1969–September 1970.

Leonid Hurwicz, University of Minnesota: visiting professor, Harvard University.

James C. Ingram, University of North Carolina: visiting professor of economics, Rockefeller Foundation, Thammasat University, Bangkok.

Paul Jonas, University of New Mexico: U.S. AID, New Delhi, India.

Jon M. Joyce, Wesleyan University: The Brookings Institution, economic policy fellowship, U.S. Department of Health, Education, and Welfare.

Ramon Knauerhase, University of Connecticut: Ford Foundation, economic advisor, Government of Saudi Arabia.

Ernest Kuhn: Fulbright lectureship, University of Liberia.

Arlyn J. Larson, Arizona State University: director of fiscal review study, legislature, state of Arizona, summer and fall, 1969.

Louis Lefeber, Brandeis University: visiting professor, Ford Foundation, Institute of Economic Growth, Delhi University.

J. C. Leith, University of Western Ontario: department of economics, University of Ghana.

Ta-Chung Liu, Cornell University: chairman, Commission on Taxation Reform of the Republic of China.

Michael C. Lovell, Carnegie-Mellon University: visiting professor, Wesleyan University.

William A. McCleary, Williams University: Rockefeller Foundation teaching and research program, Thammasat University, Bangkok.

Charles E. McLure, Jr., Rice University: senior staff economist, Council of Economic Advisors.

Andrea Maneschi, Vanderbilt University: Ford Foundation visiting professor, University of Sao Paulo.

Bruce W. Marion, Ohio State University: Harvard School of Business, Harvard University.

Stephen Marris: visiting research professor, The Brookings Institution.

H. Laurence Miller, Jr., University of Hawaii: National Bureau of Economic Research.

James N. Morgan, University of Michigan: visiting professor, Rockefeller Foundation teaching and research program, University College, Nairobi.

Masao Nakanishi, University of California, Los Angeles: visiting professor, University of Rochester.

Thomas O'Brien, Brandeis University: research director, Boston Redevelopment Authority.

Dennis J. O'Connor, Arizona State University: visiting assistant professor of economics, American University, Cairo.

Emiel W. Owens, Prairie View A and M College: visiting professor of agricultural economics, University of Minnesota.

Don Paarlberg, Purdue University: director, agricultural economics, U.S. Department of Agriculture.

George Perry, University of Minnesota: The Brookings Institution.

James R. Prescott, Iowa State University: visiting professor, University of California, Los Angeles.

Charles E. Ratliff, Jr., Davidson College: visiting professor, Forman Christian College, Lahore, Pakistan.

Paul C. Roberts, Tulane University: visiting professor of economics, University of New Mexico.

Robert L. Robertson, Jr., associate professor, Mount Holyoke College: Center for Community Health and Medical Care, Harvard University.

D. G. Sisler, Cornell University: Yale University Growth Center.

William P. Snavely, University of Connecticut: visiting professor of economics, American University of Beirut.

Charles G. Stalon, Southern Illinois University: associate professor, Economic Advisor to Federal Power Commission, Washington.

James A. Stephenson, Iowa State University: visiting professor, Special Studies Section of the Federal Reserve System, Washington.

Joseph E. Stiglitz, Yale University: Rockefeller Foundation teaching and research program, University College, Nairobi.

Herbert H. Stoevener, Oregon State University: visiting associate professor of agricultural economics, University of Minnesota.

Thomas T. Stout, Ohio State University: livestock marketing research, U.S. Department of Agriculture.

W. Paul Strassmann, Michigan State University: professor, U.N. International Labor Office, Geneva, Switzerland.

Sidney C. Sufrin, Syracuse University: visiting professor, College of Business Administration, University of Michigan.

Daniel B. Suits, University of Michigan: visiting professor, University of California, Santa Cruz.

Irving Swerdlow, Syracuse University: economic policy fellowship, The Brookings Institution.

Shanti S. Tangri, Wayne State University: visiting associate professor of economics, Duke University and senior research associate, Carolina Population Center, University of North Carolina.

Morris Teubal, University of Minnesota: Instituto Torquato Di Tella, Buenos Aires, Argentina.

Donald S. Tull, College of Business Administration, University of Oregon: Netherlands Institute of Business, Bruekelen, The Netherlands.

Martin A. Ulrich, Simon Fraser University: research associate of agricultural economics, University of Minnesota.

Raymond F. Valenti, Syracuse University: director

of graduate studies, Catholic University of Puerto Rico.

Frits van Beek, American University: staff economist, International Monetary Fund.

W. E. Vickery, University of Western Ontario: department of economics, University of Ghana.

Don J. de Voretz, Simon Frazer University: Rockefeller Foundation teaching and research program, University of Ibadan, Nigeria.

Henry A. Wadsworth, Purdue University: visiting professor, Oregon State University.

Bernard Wasow, University of British Columbia: economic consultant, Nairobi.

Eldon E. Weeks, Washington State University: Economic Research Service USDA, U.S. Department of Agriculture.

Murray L. Weidenbaum, Washington University: Assistant Secretary of the Treasury for Economic Affairs.

Finis R. Welch, Southern Methodist University: National Bureau of Economic Research.

T. Kelley White, Purdue University: Rural Institute, Purdue-Brazil Project, Vicoas, Brazil.

William E. Whitesell, Franklin and Marshall College: research department, Federal Reserve Bank of Philadelphia.

Shih-Yen Wu, University of Iowa: research professor, Belgium.

Resignations

A. M. Agapos, Ohio University: Louisiana State University.

Edward Ames, Purdue University, June 30, 1969.

Paul R. Beach, Arizona State University, June 1969.

William Beaver, Graduate School of Business, University of Chicago.

Robert L. Beck, South Dakota State University: department of agricultural economics, University of Kentucky.

E. Bell, University of British Columbia.

Michael Belshaw, Hunter College, City University of New York, Sept. 1, 1969.

John T. Berg, University of North Dakota, June 1, 1969.

Ralph E. Berry, Harvard University: School of Public Health, Harvard University.

James H. Blackman, University of North Carolina.

Martin R. Blyn, School of Commerce, New York University, Aug. 31, 1969.

Leonard Bower, Florida State University: Federal Reserve Bank, Dallas.

F. S. Chen, University of Manitoba.

George Christou, Clemson University.

Kenneth W. Clarkson, San Fernando Valley State College.

Bruce Cohen, Bernard M. Baruch College.

John E. Cooper, University of Idaho: Lewis and Clark College.

Douglas Dalrymple, University of California, January 1969.

Howard W. Davis, University of California, Riverside: Louisiana State University.

Robert D. Dean, College of Business Administration, Kent State University, June 15, 1969.

John S. DeLeeuw, Old Dominion College.

Paul Demeny, University of Michigan, Sept. 1, 1969.

Roger A. Dickinson, University of California, Berkeley.

Robert R. Dockson, Business and Graduate School of Business Administration, University of Southern California: vice chairman, Board of Directors, California Federal Savings and Loan Association, September 1969.

Don D. Doty, Mississippi State University, May 31, 1969.

George C. Eads, Harvard University, June 1969: Princeton University.

Edgar O. Edwards, Rice University: Ford Foundation, January 1970.

Eugene B. Finch, University of North Dakota, June 1, 1969.

Richard Fleming, University of Mississippi.

Richard B. Freeman, assistant professor of economics, Yale University.

William O. Freithaler, University of Virginia, August 31, 1969.

T. B. Giddings, University of Manitoba.

Frank W. Gill, University of Oregon, June 1969.

Glen A. Gransberg, University of North Dakota, June 1, 1969.

E. K. Grant, Wellington College, University of Guelph.

Sanford Gunn, School of Business Administration, University of Massachusetts.

Donald R. Hakala, Cleveland State University: Illinois State University.

Niles Hansen, University of Kentucky, July 1969.

Robert Hartman, Brandeis University, July 1969: The Brookings Institution.

Udo Heyn, University of Indiana.

Barbara L. Hodgins, University of Alberta.

C. A. Hofman, Middle Tennessee State University.

Anne E. Hulse, Hunter College, City University of New York, September 1, 1969.

Henry D. Jacoby, Harvard University, July 1, 1969: Kennedy School of Government, Harvard University.

Joe Jones, Bureau of Business and Economic Research, University of Arkansas.

Robert Jones, The City College, Aug. 1969.

John W. Judy, Jr., Purdue University, July 1969: University of Georgia.

Constantine Kalogeras, Bernard M. Baruch College.

Ira Kaminow, Ohio University: research department, Federal Reserve Bank of Philadelphia.

Michael Kawaja, University of California, Los Angeles, March 1969.

Virgil H. Ketterling, University of North Dakota, Aug. 1, 1969.

David Klahr, Graduate School of Business, University of Chicago, August 31, 1969.

John W. Leonard, College of Business and Public Administration, University of Arizona: Southern Illinois University, June 1969.

Baruch Lev, Graduate School of Business, University of Chicago, September 30, 1969.

David H. Levey, lecturer, department of economics, Yale University.

Solomon B. Levine, Institute of Labor and Industrial Relations, University of Illinois.

Robert Long, University of Mississippi.

Gerald F. McGuigan, University of British Columbia: York University.

Gilbert J. McKee, Claremont Men's College.

David C. Major, The City College, Aug. 1969.

A. Malik, University of Manitoba.

Ray Marshall, University of Kentucky, July 1969.

Arthur Matson, South Dakota State University, Sept. 1, 1969.

Harold T. Moody, University of California, Los Angeles, June 1969.

George R. Morrison, University of California, San Diego.

Tom Morton, Graduate School of Business, University of Chicago, June 30, 1969.

Bernard Munk, University of Michigan, Mar. 1969.

Paul Nadler, Graduate School of Business Administration, New York University, June 1969.

Marc L. Nerlove, professor of economics, Yale University.

Gilbert Nestel, Ohio State University, September 1969.

Corinne Norgaard, University of California, Los Angeles, June 1969.

John Owen, Johns Hopkins University, June, 1969.

William Poole, Johns Hopkins University, June, 1969.

Lee E. Preston, University of California, Berkeley.

Henry O. Pruden, University of Oregon, June, 1969: University of Texas, Austin.

Frank Puffer, University of California, Los Angeles, June 1969.

R. Raymond Radosevich, University of Kansas, July 31, 1969.

Jakob Rauber, University of Connecticut, June 15, 1969.

Kong K. Ro, assistant professor, The City College, Aug. 1969.

Ray C. Roberts, Jr., dean, School of Business Administration, Winthrop College.

Edward Rosenbaum, Lehigh University: University of Windsor, Canada.

Jeffrey M. Schaeffer, Rutgers University, June 1969.

G. Schramm, University of Manitoba.

Aba Schwartz, Ohio State University, September 1969.

Harry S. Schwartz, adviser to the Federal Home Loan Bank Board, March 1969: vice president and economist, Federal National Mortgage Association.

Gerald Scully, Ohio University: University of Southern Illinois.

Zoltan Sebestyan, Rensselaer Polytechnic Institute: New Paltz, New York.

Richard T. Seldon, Cornell University, June 1969.

Sue O. Shaw, Middle Tennessee State University.

William A. Shrode, University of Oregon, June 1969: Florida State University.

Thomas Sowell, Cornell University, June 1969.

H. Louis Stettler III, Johns Hopkins University, June 1969.

Subramanian Swamy, Harvard University, June 1969: India.

Joseph Swanson, Purdue University, June 30, 1969.

Edwin F. Terry, State University of New York, Stony Brook.

Arthur L. Thomas, University of Oregon, June 1969: McMaster University.

Wilbur Thompson, University of Michigan, Apr. 1969.

E. Hodson Thornber, Graduate School of Business, University of Chicago, June 30, 1969.

Margo Thorning, Clemson University.

Jack E. Thornton, The Citadel: Virginia Commonwealth University, Sept. 1, 1969.

Eric Trist, University of California, Los Angeles, June 1969.

David Vinje, Whitman College, June 1969.

Mei-chu Wang, University College, New York University.

Walter B. Wentz, Claremont Men's College, April 1969.

Karl A. Wesolowski, Salem State College: Raytheon, Inc.

Ned Williams, University of Mississippi.

Miscellaneous

Robert E. Gallman, University of North Carolina, Chapel Hill: editor, *The Journal of Economic History*.

Alek Rozental, National Planning Association: financial consultant, Urban Development Committee, Council for International Economic Cooperation and Development, Republic of China.